4-2021

Moral Salience, Conditional Altruism and Virtue: Reconciling Jekyll and Hyde Paradoxes

James Konow
Loyola Marymount University, jkonow@lmu.edu

FRANK COMMENTS WELCOME (NOW BEFORE YOU REFEREE IT)

April 2021

Moral Salience, Conditional Altruism and Virtue: Reconciling Jekyll and Hyde Paradoxes

James Konow
Loyola Marymount University
One LMU Drive, Suite 4200
Los Angeles, CA 90045-2659
Telephone: (310) 338-7486
Email: jkonow@lmu.edu

Abstract

The results of many observational and experimental studies reveal an economically and socially important paradox: people sometimes behave morally in certain situations but then behave immorally (or, at least, less morally) under conditions that differ for reasons that seem morally irrelevant. These patterns are inconsistent with both theories of rational self-interest as well as with theories that incorporate stable moral preferences. This paper presents a theory that reconciles many of these phenomena, including the depressing effects on moral behavior of experimentally introducing uncertainty, social distance, and options to delegate, to exit, to remain ignorant, and to take from or destroy the earnings of others. The theory introduces the concepts of moral salience and virtue preferences, which together with a model combining fairness and altruism explain not only the paradoxes but also a wide range of classic findings on distributive preferences and reciprocity. The results of an experiment that tests the theory out-of-sample proves consistent with the theoretical predictions.

Keywords: moral salience, virtue preferences, conditional altruism, fairness, altruism, reciprocity, moral wiggle room

JEL Classification: C9, D3, D9

Acknowledgements: I wish to thank Michael Berg, Prachi Jain, and Edward Mosteig for their advice, and Ben Mouehli, Gerrit Nanninga, and Eamon Shaw for their research assistance.

1. Introduction

A pedestrian gives money to beggars but, if possible, crosses the street to avoid them. Different ethnic groups live peacefully together, until genocide normalizes the destruction of life and property, as during the Bosnian War. Companies hire consulting firms to recommend or carry out the firing of their employees, although the companies could implement the firings themselves and save the consulting fees. Otherwise law-abiding citizens join in looting during civil disturbances and natural disasters. Donors in developed countries give to local causes but neglect more critical need in distant developing countries, where their support could do much more good. Some people employ uncertainty about climate change as an excuse not to act on it, even when they support measures to address less severe environmental issues. People reward or punish politicians and CEOs not only for their good or bad choices but sometimes also for uncontrollable luck. And, in the lead-up to the 2007-08 financial crisis, lenders, who were formerly prudent, chose willful ignorance and avoided documenting applicants' incomes. There are countless economically and socially important instances such as these of "Dr. Jekylls," who under certain circumstances act morally, but then, for reasons that seem morally irrelevant, behave less morally or even immorally, transmogrifying into "Mr. Hydes."

Of course, the examples above do not necessarily require an appeal to inconsistent moral preferences but might instead be explained by a variety of other factors, such as risk preferences, social image concerns, imperfect information, preemptive retaliation, strategic self-interest, fear of punishment, expectations about the behavior of others, or the benefits of specialization. Nevertheless, the results of laboratory and field experiments demonstrate that these paradoxes persist, even when one carefully controls for such factors. In an initial round that began in the 1980s (e.g., Güth, Schmittberger and Schwarze, 1982, Camerer and Thaler, 1995), experimental economists began uncovering instances of behavior at variance with the single-minded pursuit of material self-interest. In time, these initial anomalies became the "classic" results, which prompted numerous theories of stable moral (or social) preferences (e.g., see Camerer, 2003). Then, in the late 2000s, seminal experimental work, including by Dana, Weber and Kuang (2007) on "moral wiggle room," produced new "anomalies." Subjects act fairly under certain conditions but then act unfairly under slightly different conditions in ways that are inconsistent with both pure self-interest and theories that combine self-interest with stable moral preferences.

One such anomaly, studied by Bardsley (2008) and List (2007), is what I will call the

"taking effect." In the standard version of an experiment called the dictator game, one subject, called the dictator, is permitted to share part of an unearned endowment with another anonymous subject, called the recipient. Most dictator experiments show that most dictators transfer a positive amount. But when dictators are permitted not only to give but also to take from recipients' endowments in an otherwise identical treatment, many dictators take, and positive transfers also decrease in frequency. For dictators, who are fair enough to share in the standard version, the option to take in the second treatment should not matter, but it often does.

This paper introduces a theory of moral salience, conditional altruism and virtue preferences that reconciles both the classic findings on moral preferences as well as the newer anomalies that contradict both pure self-interest and stable moral preferences. It also reports the results of a new experiment that tests the theory out-of-sample and proves supportive of it. The theory concerns the preferences of an agent, who acts on a passive person, or the patient. The decision context is defined as the set of choices and information related to the choices, and it contains moral and non-moral elements that systematically affect moral salience, i.e., the prominence of moral considerations in the decision. The focus here is on the subset of moral preferences I call conditional altruism, which are allocative preferences that consist of fairness preferences and altruism, i.e., unconditional preferences to help or harm. Agents act based on the strength of their intrinsic moral preferences, which vary across agents, and the salience of those preferences, which depends on the decision context. In addition, agents are assumed to have virtue preferences: they are motivated to reward and punish others for their good or bad moral character. Moral character refers collectively to virtues, which are actions based on an intrinsic willingness to comply with different moral preferences, such as altruism and fairness.

The theory is consistent with a wide range of stylized facts, including classic findings about fairness and reciprocity as well as the anomalous results, while providing guidance about the conditions under which one can retain a social preference approach and when and how to extend it to account for anomalies. It is also related to the oldest school of thought in Western moral philosophy, virtue ethics. Along lines of this school, Ashraf and Bandiera (2017) explore how altruistic acts affect altruistic capital, and Konow and Earley (2008) discuss the relationship between virtue and happiness. The current theory relates to other features of virtue ethics, including multiple ethical principles, context-dependent morality, and a concern for virtues.

The theory is formulated around explaining and predicting non-strategic behavior in

simple experiments because of several advantages of that approach. A growing literature has demonstrated the external validity of non-strategic experiments on moral preferences quite generally, that is, pro-sociality in experiments is correlated with such behavior in the field. For example, dictator generosity is positively correlated with honesty in the field (Franzen and Pointner, 2013) and with a willingness to take costly steps to reduce the exposure of others to Covid-19 (Campos-Mercade, Meier, Schneider and Wengström, 2020). In addition, the more recent experiments on anomalies provide persuasive evidence of the internal validity of the claim that there is something inherent to moral preferences that is inconsistent with existing theories. Moral preferences are clearly relevant to important economic phenomena, such as cooperation, but cooperation is impacted by a complex set of considerations, as Dal Bó and Frechette (2018) argue. Specifically, strategic self-interest can confound inferences about morals in many contexts but should play no role in non-strategic decisions, such as in the dictator game. In particular, "virtue signaling," or feigning morally motivated behavior for strategic reasons, can distort signals about true allocative preferences as well as about virtue preferences, which are a function of the former. This accounts for the focus of this paper on non-strategic allocation decisions. Finally, simple experimental decisions enable the parallel development of a simple and tractable theory. That said, reference will occasionally be made to results from designs where strategic concerns play a potential role, in particular, with experiments where strategic concerns are likely negligible and results from non-strategic designs are not available.

A word is in order about what this paper tries, and does not try, to do. It proposes a theory of moral preferences that is novel, tractable, and capable of explaining a wide range of evidence on moral preferences, including various Jekyll and Hyde paradoxes. It reports the results of a new experiment that tests and finds support for the theory and makes some comparisons with other theories. But, for various reasons, it does not conduct a beauty contest and makes relatively few comparisons with alternative explanations. For one thing, it is already a very ambitious undertaking, a fact confirmed, in part, by its length, and further theoretical or empirical analysis is beyond the scope of a single paper. For another thing, I find many other explanations plausible in the particular cases they address, so the aim here is not to displace them. Instead, I see the chief contributions of this paper as offering a theoretical framework that is new and distinct from others, tractable, and more general in its applications to many types of behavior that are impacted by moral preferences (indeed, it might be seen as a generalization of some prior theories).

Section 2 presents the theory in general terms. Section 3 discusses some general applications of allocative preferences and section 4 some general applications of virtue preferences. The two sections thereafter address classes of anomalies: section 5 helping and harming, including the new experiment, and section 6 norm avoidance. Section 7 discusses briefly a different type of moral salience called point salience, and section 8 concludes.

2. Theory

This section introduces a theory of moral salience, integrates salience into and generalizes a model of allocative preferences called conditional altruism, and presents a theory of preferences to reward and punish called virtue preferences.

2.1. Moral Salience

Economists have cited salience in connection with various phenomena, including consumer choice (Bordalo, Gennaioli, and Shleifer, 2013, 2016), strategic decision-making (Crawford, Gneezy and Rottenstreich, 2008, Crawford and Iriberri, 2007), taxation (Chetty, Looney, and Kroft, 2009), and the endowment effect (Bordalo, Gennaeioli, and Shleifer, 2012). This paper also presents a theory of salience as prominence but, like Benabou and Tirole (2006), in a specifically moral context. It is, however, distinct from prior formalizations, to my knowledge. *Moral salience* concerns how the decision context affects the prominence of moral considerations in individual choices. This subsection introduces a concept of salience that characterizes how properties of subsets of the context affect the prominence of moral preferences. This is the primary concept of moral salience employed in this paper, but a different type of moral salience will be discussed briefly in section 7 of the paper, which involves the prominence of specific actions within the set of available actions.

Consider a person, called the *agent*, who makes a decision that materially affects a passive individual, called the *patient*.[1] This might be a sponsor choosing how much to donate to a child supported by a charitable organization or a dictator deciding how much of an endowment to transfer to a recipient in a dictator game. The agent may take an action, x, from the set of available actions, X. In most of the situations considered in this paper, the action is the same as

[1] Lacking a general and commonly agreed upon term in economics for a person who is acted upon by a moral agent, I borrow the term patient from philosophical ethics.

the material effect on the patient, e.g., the transfer received by the recipient in a dictator game, which is selected from the range of permissible transfers, i.e., $x \in X$. The agent also possesses information that might be seen as morally relevant to the choice of actions. For example, a dictator might be informed that the recipient has an endowment, *y*, among other elements of the set of information about the decision, i.e., $y \in Y$. Such actions and information are elements, *c*, of the decision context, *C*, i.e., $c \in C$, whereby $C = X \cup Y$.

Moral set salience is the weight attached to the agent's moral preferences as a result of the decision context. For example, the taking effect described in the Introduction is consistent with the interpretation that adding taking options to the set of available actions in a dictator game reduces the weight on moral preferences and, therefore, the level of dictator transfers. Similarly, information indicating greater social distance between dictator and recipient can lower the weight on the agent's moral preferences, as discussed later.

We will consider in future sections various ways in which qualitative elements of context can affect moral salience. But let us first restrict attention to quantitative contextual elements, such as giving and taking options in a dictator game:

$C = \{c\}, c \in \mathbb{R}$.

Let *C* be partitioned into those elements in the moral context, C_+, and those in the non-moral context, C_-:

$C = \{C_+, C_-\}$

The elements of the former are positively valenced and increase moral salience, e.g., opportunities to help another person. The latter comprise those elements that are negatively valenced and diminish moral salience, e.g., opportunities to harm another. Non-moral context can also include amoral, or morally neutral, elements, e.g., the possibility of inaction. Define a function, $m(C_i)$, of these partitions, $C_i = \{C_+, C_-\}$, that satisfies the properties of a measure, viz., non-negativity ($m(C_i) \geq 0 \forall C_i \in C$), null empty set ($m(\emptyset) = 0$), and countable additivity ($m(\bigcup_i C_i) = \sum_i m(C_i)$). Denote the moral measure, $p = m(C_+)$, and the non-moral measure, $n = m(C_-)$. Distinguishing moral and non-moral context and constructing measures of them requires, of course, some judgment and depends on the decision context. In the various applications that follow, we discuss different commonsensical specifications for these measures.

Now we come to moral salience, which is related to the usual understanding of salience in neuroscience and social psychology. In those disciplines, salience typically refers to how an

object, or set of objects, stands out relative to its environment. Here, moral salience refers to how subsets of elements of the decision context affect the prominence of moral considerations and, therefore, the weight on an agent's moral preferences. This is a version of salience I call "set salience," and it involves collections of objects that are all disjoint subsets of a superset. The elements of each subset share some feature(s) (here, whether they are moral or non-moral), and each subset distinguishes itself in this way from other subsets. I focus here on the case in which the context may be bifurcated into measurable subsets. Set salience refers to the tendency for the subset with smaller measure to have disproportionate prominence relative to the contrasting subset with larger measure. For example, a five-year-old does not stand out in a Kindergarten but does in a retirement home. The set salience I propose further specifies a non-linear relationship between salience and measures of the subsets of context. Returning to our anthropomorphic example, a comparatively small group of children situated among older people is prominent, but the marginal salience of the first child is greater than that of the second and the marginal salience of the second is greater than that the third, etc. Moral set salience formalizes this property for moral preferences. The addition of moral context to a mostly (or entirely) non-moral context, and the related increase in the moral measure, increases moral salience and does so at a decreasing rate. Conversely, the addition of elements of non-moral context to a mostly (or entirely) moral context, and the attendant increase in the non-moral measure, decreases moral salience at a decreasing rate, that is, the first addition non-moral context causes a larger decrease in the prominence of moral considerations than the second, etc.

Formally, consider the following definition, which reflects these properties.

DEFINITION 1: Moral set salience, $\sigma(p, n)$, is a function with support on the non-negative real numbers that maps the moral and non-moral measures of the decision context into the unit interval:

$$\sigma: \mathbb{R}_+^2 \to [0,1].$$

It is assumed that $\sigma(p, 0) > 0, p > 0;\ \sigma(0, n) \geq 0, n > 0$; and that $\sigma(p, n)$ is twice continuously differentiable with

$$\frac{\partial \sigma}{\partial p}\Big|_{n>0} > 0, \frac{\partial^2 \sigma}{\partial p^2}\Big|_{n>0} < 0, \frac{\partial \sigma}{\partial n}\Big|_{p>0} < 0, \frac{\partial^2 \sigma}{\partial n^2}\Big|_{p>0} > 0.$$

It proves convenient in the subsequent analysis to flesh out this function in a more specific form. An expression that captures the assumed relationships between σ and p and n is

the ratio $\frac{p}{p+n}$. Many decisions, however, involve some fixed moral salience with variation in only a subset of the moral context. For example, in a dictator game, variation in the amounts one may transfer might impact moral salience through its effects on the measures *p* or *n*, but there are often some baseline moral considerations, e.g., triggered by the very fact of being endowed and paired with another person. In such cases, the context contains a baseline, or fixed, moral set salience denoted $\bar{\sigma} \in [0,1)$ in addition to the subsets of moral context that are variable, *p* and *n*. This leads to the following specification for moral salience:

(1) $\sigma = (1 - \bar{\sigma}) \cdot \frac{p}{p+n} + \bar{\sigma}.$

This expression satisfies the conditions that define moral salience: its maximum is $\sigma(p, 0) = (1 - \bar{\sigma}) \cdot \frac{p}{p} + \bar{\sigma} = 1$, its minimum is $\sigma(0, n) = \frac{0}{n} = 0$ when $\bar{\sigma} = 0$, $\frac{\partial \sigma}{\partial p} = (1 - \bar{\sigma}) \frac{n}{(p+n)^2} > 0$, $\frac{\partial^2 \sigma}{\partial p^2} = -(1 - \bar{\sigma}) \frac{2n}{(p+n)^3} < 0$, $\frac{\partial \sigma}{\partial n} = -(1 - \bar{\sigma}) \frac{p}{(p+n)^2} < 0$, and $\frac{\partial^2 \sigma}{\partial n^2} = (1 - \bar{\sigma}) \frac{2p}{(p+n)^2} > 0$.

Figure 1 illustrates how moral salience varies with the moral and non-moral measures. The aforementioned effects of moral and non-moral context on moral salience are reflected in the properties of σ being increasing and concave in *p* for a given $\bar{n}$ and decreasing and convex in *n* for a given $\bar{p}$. Fixed moral salience, $\bar{\sigma}$, represents the lower bound of moral salience.

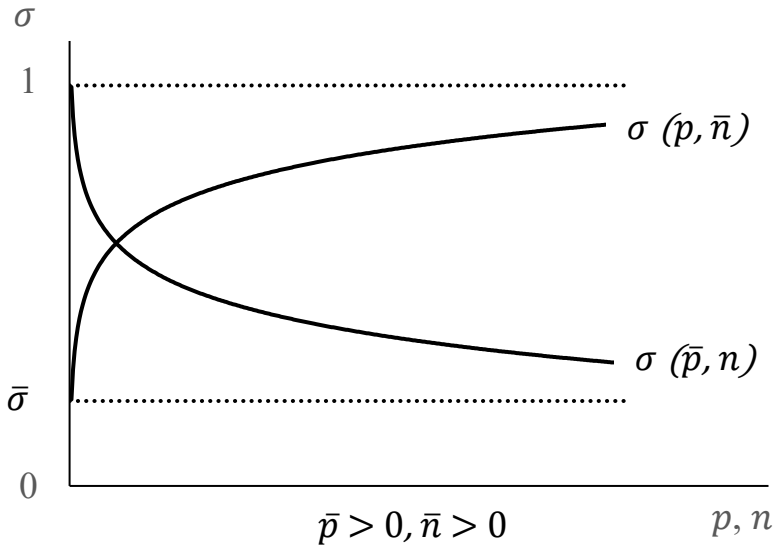

FIGURE 1. – Moral salience.

The remaining sections analyze numerous contextual factors that affect moral salience. Some cases involve binary decisions, such as whether or not to remain ignorant of information that raises a moral obligation. In other cases, however, there is empirical evidence on the effects

of incremental changes in moral or non-moral context. For instance, moral salience may vary with the amounts that may be given to or taken from a patient, physical proximity to the patient, and probability that the agent's decision is actualized. These cases lend themselves to cardinal measurement, so that one can observe not only the direction of the effect of context on moral salience but also differences in the rate of change in that effect.

Another practical aspect is that people are sometimes confronted with multiple decisions in similar moral contexts at the same time. This occurs, for example, in experiments that present the same group of subjects with similar decisions in a within-subjects design. It also arises, though, outside the laboratory, e.g., when someone receives multiple solicitations to donate to different charities. In such cases, I make the following assumption.

ASSUMPTION 1: Denote the contexts of decisions 1 and 2, respectively, C^1 and C^2. Suppose they are related, meaning choices are made jointly from decision contexts that are identical except for some element, c: $\{C^1/c^1\} = \{C^2/c^2\}$ and $c^1 \neq c^2$. Then the decisions share the common context $C = \{C^1 \cup C^2\}$ with the same measures and the same moral salience.

The examples stated thus far indicate that moral set salience is affected by a variety of contextual factors, which differ qualitatively. These factors are worked out in further sections of the paper, but in order to sort through these factors, it is helpful to clarify in general terms how to distinguish what properties of the context affect moral set salience and in what way. I approach this by identifying the properties that make for the highest moral salience and arguing that relaxing these lowers moral salience. It is based on the following "empirical" assumption, i.e., an assumption about the estimation or empirical interpretation of a theoretical construct.

ASSUMPTION 2: Moral salience is highest when there is a single agent and single patient, who stand in the close proximity. In addition, the choices open to the agent comprise only non-harming actions, the effects of the actions on the patient are certain and common knowledge, moral norms relevant to the context are explicit and common knowledge, and there are no opportunities to avoid actions or information about the consequences of actions that relate to moral norms.

Thus, the focus of the analysis of moral salience in this paper is on the ways in which non-moral context may reduce salience. These include by adding context that reduces proximity, adds harmful actions to the choice set, or provides self-serving opportunities to avoid or delegate the decision or to avoid information about the consequences of the decision. Note that, since the

moral context includes the information provided, *Y*, moral salience can be subject to framing effects, e.g., a dictator's transfer can be affected by whether the task is worded as a choice to "give" or to "distribute" money.[2] Such changes in presentation might or might not affect choices, depending on whether they affect moral salience. But the effects on moral salience are not limited to framing effects, since moral salience is also a function of the actual set of available choices, *X*, and not just their presentation, e.g., whether a dictator may take as well as give. Thus, context can affect choices mediated by salience even under perfect information due to differences in choice sets.

2.2. Conditional Altruism with Moral Salience

As already stated, the theory presented here seeks to offer a unified framework that is consistent with both classic and anomalous results on moral preferences. This theory weights moral preferences by moral salience, but a critical question concerns the type of moral preferences. As previously stated, this paper focuses on allocative preferences, specifically, the model of *conditional altruism* introduced in Konow (2010). This section has several goals. It extends this model, generalizing the altruism term, elaborates new implications of the model, and analyzes some implications of integrating moral salience into it.

Conditional altruism has three components: material utility and two moral motives, fairness and altruism. Material utility, $u: \mathbb{R}_+ \rightarrow \mathbb{R}_+$, is assumed to be a twice continuously differentiable function of the agent's material allocation, π_i. Specifically, I assume material utility, $u(\pi_i)$, is as follows

$u(0) = 0, u'(\pi_i) > 0$, and $u''(\pi_i) \leq 0$.

The "conditional" part of conditional altruism involves fairness, which refers here to moral preferences over the patient's allocation that are conditioned on any distributive norm, whether equality, equity, efficiency, need or something else. I assume that the relevant moral norm depends on the context. The behavior of stakeholders, however, such as dictators in dictator games or proposers in ultimatum games, typically reflects not only moral preferences but also self-interest. So, I propose that the moral norm, which I will call the *entitlement*, be inferred empirically from the behavior of third-party allocators, as stated in the following assumption.

[2] See Bergh and Wichardt (2018) for evidence of the effects of this wording on dictator transfers. In fact, information might even be false but relevant to the agent's actions, if it influences moral salience, although providing false information is typically taboo in economics experiments.

ASSUMPTION 3: The entitlement, denoted η, refers to the moral allocation or allocative rule that is the governing moral norm of stakeholders in a given context, *C*. The entitlement can be inferred from the allocations of spectators, or third-party allocators, in *C*.

This method for eliciting impartial moral views was introduced in Konow (2000) and has been employed and elaborated in numerous subsequent studies, including Aguiar, Becker and Miller (2013), Almås, Cappelen and Tungodden (2020), Cappelen, Konow, Sørensen and Tungodden (2013), Croson and Konow (2009), Konow (2005, 2012), Konow, Saijo and Akai (2020), and Møllerstrom, Reme and Sørensen (2015). Various studies compare stakeholder and spectator decisions and reveal properties that bolster the case for employing spectator allocations to gauge the role of moral norms on stakeholder behavior. Stakeholder decisions correlate significantly with the entitlements inferred from spectators, and the variance around spectator decisions is significantly lower than that around stakeholder decisions, as one would expect, if stakeholders differ in their relative degree of self-interest. Thus, I will sometimes refer in this paper to evidence from spectator views, both in prior studies as well as in the new experiments reported here, to identify the entitlement that is relevant in a given context.

Here fairness is based on one or more values or rules from a set of possible moral norms and not only the more narrow understanding of fairness as equality or equity but also, for example, possibly efficiency or need. This is expressed as a function, $f: \mathbb{R}^2 \rightarrow \mathbb{R}_{\leq 0}$, that captures the preference of the agent, *i*, over the material allocation, π_j, of the patient, *j*, relative to the patient's entitlement, η_j. Specifically, I assume *f* is the twice continuously differentiable function

$$f\big(\phi_i(\pi_j - \eta_j)\big),$$

where $f(0) = 0, f'(\phi_i w) \cdot w < 0$ for $w = \pi_j - \eta_j \neq 0$ and $\phi_i > 0$, and $f''(w) < 0$. Agents are assumed to differ in the strength of their fairness preference, which is captured by the fairness coefficient $\phi_i \in \mathbb{R}_+$. This coefficient is distributed according to the cumulative distribution function $\Phi(\phi_i)$, where $\Phi(\phi_i)$ has support $[\underline{\phi}, \overline{\phi}]$ with $0 < \underline{\phi} < \overline{\phi} < \infty$ and $0 < \Phi\left(\underline{\phi}\right) < 0.5$. The assumptions about $\underline{\phi}$ help establish predictions that are consistent with behavior discussed later, viz., that all agents care somewhat about fairness and that minimally fair types constitute a minority. Furthermore, fairness is assumed to be homogeneous of degree

λ, such that this term may also be written $\phi_i{}^{\lambda} f(\pi_j - \eta_j)$.[3] Note that agents experience disutility, when patients have more or less than their entitlement. That is, fairness is never utility increasing, which reflects the idea that moral norms signify an obligation rather than an opportunity. This is a critical factor for later explaining a number of empirical findings.

The "altruism" part of conditional altruism refers to a moral preference that is personal and unconditional. As with standard theories of altruism, it is not conditioned on a moral norm, such as equality or efficiency, or on the behavior of others, such as a desire to reward or punish deviations from norms. Here I generalize the prior version of this model, which formally resembled warm glow (e.g., Andreoni, 1989), to include explicitly not only positive but also negative altruism (i.e., spite), as in Levine (1998). Unlike pure altruism but like warm glow, it is assumed to be a function solely of that part of the patient's allocation that can be attributed to a personal choice of the agent, e.g., a dictator making a transfer to or from a recipient in a standard dictator game. Unlike pure altruism, it is not a function of the patient's total allocation or of any amounts the patient receives from others. Altruism is also personal in that it assumed to apply to agent-patient relationships but not to impartial third party, or spectator, decisions. Altruism is expressed as a function, $g: \mathbb{R} \to \mathbb{R}$, of the amount, x_j, the patient receives from the agent.

Specifically, I assume altruism is the twice continuously differentiable function

$g(\alpha_i \cdot x_j)$

where $g(0) = 0, g'(\alpha_i \cdot x_j) > 0$ for $\alpha_i > 0, g'(\alpha_i \cdot x_j) < 0$ for $\alpha_i < 0$, $g'(\alpha_i \cdot x_j) = 0$ for $\alpha_i = 0$, and $g''(\alpha_i x) < 0$ for $\alpha_i \neq 0$. Agents differ according to their altruism coefficient, $\alpha_i \in \mathbb{R}$, and are categorized as altruistic, $\alpha_i > 0$, selfish, $\alpha_i = 0$, or spiteful, $\alpha_i < 0$. The altruism coefficient is distributed according to the cumulative distribution function $\mathrm{A}(\alpha_i)$, which has support $[\underline{\alpha}, \overline{\alpha}]$ with $-\infty < \underline{\alpha} < 0 < \alpha^* < \overline{\alpha} < \infty$, where $\alpha^* = \{\alpha | u'(X - \eta) = \sigma \alpha g'(\alpha \eta)\}$ where σ denotes the level of salience in the standard dictator game. I assume $0 < \mathrm{A}(0) < 0.5$, $0 < \mathrm{A}(\overline{\alpha}) - \mathrm{A}(\alpha^*) < 0.5$, $\mathrm{A}(\overline{\alpha}) - \mathrm{A}(0) > 0.5$, and $\int_{\underline{\alpha}}^{\overline{\alpha}} \alpha_i \, \rho(\alpha_i) d\alpha_i > 0$, where $\rho(\alpha_i)$ is the probability density function of α_i. That is, I assume a minority of agents is spiteful, a majority is altruistic, and the average type is altruistic. In addition, a minority in the standard dictator game is so altruistic ($\alpha_i > \alpha^*$) that their marginal altruism exceeds their marginal material utility evaluated where

[3] This homogeneity assumption later proves convenient for identifying threshold values of fairness for different categories of behavior, when the choice space is discrete.

their transfer equals the patient's entitlement. The altruism term accommodates positive transfers to the patient, $x_j > 0$, as well as negative ones, $x_j < 0$, i.e., taking from the patient. Note that a given x_j can increase or decrease an agent's utility. For example, for an altruistic agent, giving increases utility and taking reduces utility. The giving case is similar to warm glow, but this term additionally incorporates disutility from taking. Note that this term is upward sloping for $\alpha_i > 0$ and downward sloping for $\alpha_i < 0$, i.e., the utility of a spiteful agent decreases with giving and rises with taking. Finally, I also assume that ϕ and α are not negatively correlated. That is, although there can be exceptions, fairer agents are not, on average, less altruistic. This seems like a reasonable assumption, and it proves useful later when dealing with virtue preferences.

I assume additively separable utility, keeping with most social preference models, e.g., Charness and Rabin (2002), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), Fehr and Schmidt (1999), and Rabin (1993). Letting the moral preference terms be weighted by moral salience, σ, the utility of agent *i*, U_i, becomes

(2) $$U_i = u(\pi_i) + \sigma \cdot f\big(\phi_i(\pi_j - \eta_j)\big) + \sigma \cdot g\big(\alpha_i \cdot x_j\big).$$

There are sometimes arguments for separate moral salience variables for different moral preferences, i.e., for distinguishing fairness salience, σ_f, from altruism salience, σ_g. Nevertheless, the results almost never depend on such independent variation, indeed, only one part of one theorem, viz., the last claim in Theorem 5.3.1, hinges on such a difference, so I simplify the analysis and use a single moral salience term. One question that must be addressed, though, is the treatment of salience in the case of spiteful agents. For spiteful agents, the salience variable that is applied to their altruism term will be defined as $\sigma_b \equiv 1 - \sigma_g$, where $\sigma_b \in [0,1]$. This has an intuitive implication: a context that, for altruistic agents, increases the utility (disutility) from giving (taking), for spiteful agents, reduces the utility (disutility) from taking (giving). Finally, when uncertainty is involved, I assume that decision-makers are expected utility maximizers.

As previously stated, the theory is formulated for non-strategic decisions, and much of the focus is on the dictator game and variations on it. So, I adapt equation (2) to this experiment and simplify some notation to represent the utility of the dictator. Let *X* represent the endowment of the dictator and *x* the dictator's transfer to the recipient such that $\pi_i = X - x$. The recipient's endowment, in those cases where it is relevant, is denoted *Y* such that $\pi_j = Y + x$. Then, for the

dictator game, equation (3) can be written (suppressing subscripts for individuals)

(3) $U = u(X - x) + \sigma \cdot f(\phi(Y + x - \eta)) + \sigma \cdot g(\alpha x)$.

Now we turn to some theorems about transfers in the dictator game, which will come in handy in the later analysis. All refer to the mean dictator and, therefore, involve interior solutions, and their proofs appear in the Appendix.

I begin with the effects of moral salience.

THEOREM 2.2.1: The optimal transfer, *x*, is increasing in σ.

As stated above in the section 2.1, the analysis proceeds from a reference point of high moral salience and focuses on the effects of non-moral context, *n*, on reducing salience. So, it proves useful to establish the relationship between *n* and x, which is addressed in Theorem 2.2.2.

THEOREM 2.2.2: The optimal transfer, *x*, is decreasing in *n*. Assuming *x* is weakly convex in σ, *x* is strictly convex in *n*.

This theorem states that non-moral context decreases giving due to the reduction in moral salience. In addition, it establishes that, in the case of cardinal measures of *n*, giving decreases at a decreasing rate, assuming *x* is weakly convex in σ.[4] That is, the initial addition of non-moral context causes a larger decrease in giving than the next.

For two reasons, the remainder of this paper focuses on the effects of variation in *n* rather than *p*. First, theory yields a stronger prediction about *n*: from Theorem 2.2.2, $\frac{\partial^2 x}{\partial n^2} > 0$, but the sign of $\frac{\partial^2 x}{\partial p^2}$ is ambiguous due to the concavity of σ in p. Second, most of the experimental evidence related to moral salience seems more directly related to variation in n. Moreover, the few studies of which I am aware that explicitly relate to variation in p are consistent with the theoretical claim that $\frac{\partial x}{\partial p} > 0$ but do not involve cardinal measures needed to shed light on the second derivative, e.g., dictator giving increases significantly with the addition of a short

[4] Note that the assumed relationship between *x* and σ is a feature of several commonly used parametric utility functions. For example, for the standard dictator game, suppose we can write $U = X - x + \sigma h$, where $h = f + g$. This formulation treats material utility as linear in the dictator's payoff, as commonly assumed in many social preference theories, e.g., Charness and Rabin, 2002, Dufwenberg and Kirchsteiger, 2004, Falk and Fischbacher, 2006, Fehr and Schmidt, 1999, and Rabin, 1993, and as a special case of the current theory (since $u' > 0$ and $u'' \leq 0$). This assumption seems innocuous for economics experiments, where the stakes are usually modest relative to subjects' income or wealth. Then, it is straightforward to show that, if $h = a + b \cdot \ln x, a, b > 0$, then $dx/d\sigma = b > 0$ and $d^2x/d\sigma^2 = 0$, and, if $h = a + b \cdot x^{1/2}, a, b > 0$, then $dx/d\sigma = \frac{1}{2}b^2\sigma > 0$ and $d^2x/d\sigma^2 = \frac{1}{2}b^2 > 0$, both of which are consistent with (weak) convexity of *x* in σ.

statement about the recipient's reliance on the dictator (Brañas-Garza, 2007), and trading volume in an experimental market decreases when a negative externality is added (Sutter et al., 2020).

Finally, consider the effects on transfers of changes in ϕ, α, and η.

THEOREM 2.2.3: The optimal transfer is increasing in the fairness coefficient, ϕ, except for super-fair dictators, for whom it is decreasing in ϕ.

THEOREM 2.2.4: The optimal transfer is increasing in α.

THEOREM 2.2.5: The optimal transfer is increasing in the entitlement, specifically, $0 < {dx}/{d\eta} < 1$.

Thus, stronger moral preferences result in higher transfers, except for super-fair dictators, who experience increased disadvantageous inequity aversion. Theorem 2.2.5 implies that a one unit increase in the entitlement produces a less than one unit increase in the transfer, which is due to diminishing marginal altruism and the increasing marginal material disutility.

2.3 Virtue Preferences

The utility function presented thus far is entirely consequentialist, i.e., it represents preferences over outcomes or the consequences of decisions. Material utility reflects the material allocation of the agent. Conditional altruism, captured by fairness and altruism, are allocative preferences with respect to the fair transfer and the agent's endowment, respectively. This section presents an additional moral motive, viz., to *sanction*, that is, to reward or punish others.[5] Although we do not explore this aspect here, in a dynamic framework, such preferences might serve to undergird virtue, indeed, Dal Bó and Dal Bó (2014) find that punishment is a critical force that sustains the favorable effects of increased moral salience on cooperation.

Most theoretical accounts of sanctioning are *reciprocity* theories, e.g., Charness and Rabin (2002), Falk and Fischbacher (2006), Rabin (1993), and Dufwenberg and Kirchsteiger (2004). These theories formulate a motive to reward or punish others based on their so-called intentions. Good or bad intentions are inferred based on both others' (expected) choices and their available choice set. For example, one formulation considers whether the expected consequences of another's action exceed or fall short of some "fair" benchmark, whereby the benchmark is

[5] In the absence of any other generally agreed upon term, I use the term sanction collectively for both reward and punishment, since its meanings include both kind and unkind actions, viz., in verb form, "to give approval to" and "to impose a penalty upon" (Merriam-Webster.com).

defined relative to the other's available choice set. An alternative approach is to formulate this motive with respect to *moral types*. Levine (1998) introduced such a model in which the sanctioning motive depends jointly on the altruism (or spite) of both the agent and the patient. In his model, an agent might be altruistic or spiteful but is more altruistic (spiteful) toward a more altruistic (spiteful) patient.

This paper introduces a theory of sanctioning I will call *virtue preferences*. This represents an interpretation of virtue ethics, which is the oldest school in Western moral philosophy. Its advocates span centuries and include Aristotle (1925), Adam Smith (1759), and Martha Nussbaum and Amartya Sen (1993). It should be noted that there are different schools of thought within virtue ethics, but the theoretical concepts presented here are variations on common positions that can be found in that branch of ethics. Morality is viewed as pluralistic, i.e., consisting of multiple virtues, whereby in the present case of allocative preferences, these virtues constitute fairness and altruism.[6] As interpreted in the present theory, a virtue is a specific type of willingness to act morally, e.g., to be fair or altruistic, that is actually acted on. This differs from the approaches reviewed above in subtle but important ways. Unlike intentions in reciprocity theories, virtue is not a consequence (or intended consequence), because it is a function of moral motivation. Virtue is closer to the moral types approach, given its link to moral preferences, but it is also distinct in several ways. First, moral preferences must be realized in action, and, unlike the moral types approach, the relevant metric is behavioral and not the latent variable of moral preferences. Second, and in a related point, the sanctioning motive in virtue preferences is not over the moral preferences of others but over their actions. For example, in a dictator game, fair dictators are not rewarded for having fair preferences per se but for behavior that manifests their fairness preferences. Neither intent alone nor action alone suffices: rather, virtue is intent (understood as moral preferences) coupled with action. Finally, virtue is the *notional* willingness to behave morally, even if the *effective*, or actual, behavior differs due to obstacles that preclude more precise expression of that willingness because choice is subject to constraints or uncertainty. For example, a dictator, who is willing to share $12 with a recipient is more virtuous than a dictator, who is only willing to share $10, even both share the same $10 due to experimental rules that cap transfers at $10. Note, however, that, to count as virtue, action

[6] In the standard Aristotelian terminology, these two virtues correspond to justice and liberality, respectively, although Aristotle also discussed other virtues, among them prudence, courage, truthfulness, and friendship.

must be involved, even if the effective action differs from the notional one.

Moral character refers to an individual's set of virtues. Of course, depending on the context, multiple virtues might be in play. For example, in the contexts examined here, conditional altruism predicts that moral behavior reflects both fairness and altruism. Virtue ethicists argue that the relative importance of the different virtues in determining right action is context-dependent. The theory introduced here proposes the form of this context-dependence: moral and non-moral context determine the relative and absolute salience of the virtues. In the context of allocative preferences, moral character reduces to intrinsically motivated *generosity*, that is, the willingness to sacrifice materially, whether for the sake of fairness, altruism or both. Denote this γ, where $\gamma \in \mathbb{R}$, and suppose γ is distributed according to the cumulative distribution function $\Gamma(\gamma)$, where $\Gamma(\gamma)$ has support $[\underline{\gamma}, \overline{\gamma}]$ with $-\infty < \underline{\gamma} \leq 0 < \overline{\gamma} < \infty$. This is the variable that individuals are assumed to be motivated to sanction.[7] Specifically, the morally relevant generosity is notional, so I will conceptualize it with the following *reference state*. Consider an agent, who may make a unilateral, anonymous and unlimited transfer of resources to or from a patient. There is no role for strategic self-interest, and the agent's only motivation for departing from narrow material interests is moral preferences. The reference state is like a standard dictator game, except that the dictator has unlimited access to the entire material endowments of both parties. Further assume that, in this hypothetical dictator game, even the most generous dictator would keep some his or her own endowment ($\overline{\gamma} < X$) and that even the most selfish dictator would not take all of the recipient's endowment ($\left|\underline{\gamma}\right| < Y$). Although not necessary, this assumption simplifies the analysis, but it also seems plausible, at least for populations akin to those that participate in economics experiments and who would likely neither give away their last material possession nor take away the last possession of another.

Virtue preferences are preferences over moral character, specifically, preferences to reward the good, or praiseworthy, moral character of another, or to punish the bad, or blameworthy, moral character of another. In the present case, moral character is an expression of allocative preferences, which reduces to generosity, so that virtue preferences are preferences

[7] Note that the moral types formulation produces a counterintuitive implication for fairness preferences with super-fair agents: in this case, a fairer agent is less generous (Theorem 2.2.3) but more deserving of reward. Formulating the variable that is sanctioned in behavioral terms, in which only disadvantageous inequity aversion contributes to generosity, avoids this implication.

over others' generosity. These sanctioning preferences consist of several parts. Let us begin with an agent, denoted *k*, who is capable of sanctioning others. This person may, depending on the decision context, be stakeholder or a third party. This agent may transfer an amount to or from another, denoted $z_k \in \mathbb{R}$, whereby the range of possibilities considered in this paper comprises $z_k \in [\underline{z}, Z]$, where $-\infty < \underline{z} \leq 0$ and $0 \leq Z < \infty$. The agent may make transfers for allocative reasons but also in order to sanction, i.e., to reward or punish another beyond what allocative preferences alone demand. The agent's ideal level of sanctioning is denoted $\tilde{z}_k$, and this depends on *k*'s estimate of the moral character (in the present case, *k*'s estimate of the sanctioned person *j*'s notional generosity), $\hat{\gamma}_j$. Since the decision context differs from the thought experiment described above, *j*'s actual moral character, γ_j, is not known and must be estimated. As explained in later sections, $\hat{\gamma}_j$ is based on *j*'s choices as well as the decision context, including the reigning moral salience and constraints on *j*'s choices. In addition, for each agent *k* and given the level of salience in the reference state, $\tilde{\sigma}$, there is a threshold value of γ_j, denoted $\tilde{\gamma}_k$, where $\underline{\gamma} < \tilde{\gamma}_k < \overline{\gamma}$. Above this threshold, agent *k* judges *j*'s character as praiseworthy and deserving of reward and, below it *j*'s, character is viewed as blameworthy and deserving of punishment. This "character threshold" may differ across sanctioning agents.

The ideal sanction, $\tilde{z}_k$, is assumed to depend on $\tilde{\gamma}_k$ and $\hat{\gamma}_j$ through the term $r: \mathbb{R} \to \mathbb{R}$, which is the twice continuously differentiable function

$$r\left(\hat{\gamma}_j - \tilde{\gamma}_k\right)$$

where $r(0) = 0, r'\left(\hat{\gamma}_j - \tilde{\gamma}_k\right) > 0$, and $r''\left(\hat{\gamma}_j - \tilde{\gamma}_k\right) < 0$.

This implies a positive ideal sanction, $\tilde{z}_k > 0$, when estimated character exceeds the threshold, $\hat{\gamma}_j > \tilde{\gamma}_k$, and a negative $\tilde{z}_k < 0$, when the opposite is the case, $\hat{\gamma}_j < \tilde{\gamma}_k$. In addition, the concavity of r captures the idea that blameworthy character implies greater punishment than the reward for praiseworthy character of an equal degree. I assume that agents *k* care in differing degrees about sanctioning, denoted $\theta_k \geq 0$, which is distributed according to the cumulative distribution function $\Theta(\theta_k)$, where $\Theta(\theta_k)$ has support $[\underline{\theta}, \overline{\theta}]$ with $0 = \underline{\theta} < \overline{\theta} < \infty$, and $\Theta(0) > 0$. That is, there is a mass of people, who care not a whit about sanctioning. To r and θ_k we add the scale of the sanction, x', which specifies the magnitude of reward or punishment appropriate to the context. This provides a measure of the importance of the action, and to ignore the scale of the decision context, when choosing how to sanction character, would have implausible

implications, such as taking a bad person, who is curt with a waiter, to be equally deserving of punishment as someone of equally bad character, who robs a bank. Since virtue is willingness coupled with action, and people are not sanctioned merely for being, character must be matched with a context that involves choice. In contexts with certainty about the mapping of choices to allocations, I propose defining the scale as the transfer called for by the moral norm, viz., the patient's entitlement such that $x' = \eta_j$ (I will present a more general specification that includes uncertainty later). Thus, the ideal sanction can be expressed

$$\tilde{z}_k = \theta_k \cdot r\left(\hat{\gamma}_j - \tilde{\gamma}_k\right) \cdot x'.$$

Finally, to sanction means to increase or decrease the patient's payoff beyond what is called for by distributive preferences alone (i.e., η_j), so $\tilde{z}_k$ is incorporated into the fairness function. The complete specification of the utility function of an agent with material utility, fairness, altruism and virtue preferences is

$$U_i = u(\pi_i) + \sigma \cdot f\left(\phi_i\left(z_k - \eta_j - \tilde{z}_k\right)\right) + \sigma \cdot g(\alpha_i \cdot z_k). \quad (6)$$

Various specifications of this utility function will be fleshed out in the sections that follow. Note that subscripts are suppressed, wherever subject roles are clear.

3. Applications of Allocative Preferences

One of the main goals of this paper is to reconcile stylized facts from a wide range of prior experiments with the theoretical framework advanced here. This section discusses applications of the allocative theory with moral salience to findings that frequently emerge across various contexts, including both classic and anomalous results.

3.1. Classic Results

This section demonstrates the consistency of the theory with classic results of social preference experiments. Below is a list of stylized facts (SF), each of which is accompanied by (and sometimes identical to) a theorem that asserts a claim about the consistency of the SF, or parts of it, with the theory. This is followed by proofs, which are based on the theory, explicitly stated additional assumptions and/or previously stated stylized facts. Note that parts of some SF are simply taken as empirical regularities and are not proven, and some may be used in subsequent proofs. Minor or lengthier proofs are relegated to the Appendix.

SF 3.1.1: There is a mass at null transfers in the standard dictator game (e.g., 36%, on average,

across multiple studies in the survey of Engel, 2011).

THEOREM 3.1.1: Suppose for the least fair dictators (i.e., those with $\underline{\phi}$) in the standard dictator game with salience denoted σ it is the case that $u'(X) > \sigma\underline{\phi}f'\left(-\underline{\phi}\eta\right)$. Then there is a mass at null transfers in the standard dictator game.

PROOF: In the standard dictator game, transfers are constrained to be non-negative, so a corner solution results at $x = 0$ among that fraction of dictators who are comparatively self-interested, i.e., for whom $u'(X) \geq \sigma\phi f'(-\phi\eta) + \sigma\alpha g'(0)$. Specifically, sufficient conditions for this are the mass of dictators, who are both not altruistic (i.e., $\mathrm{A}(0) > 0$), and who are also the least fair such that at $u'(X) > \sigma\underline{\phi}f'\left(-\underline{\phi}\eta\right)$.

SF 3.1.2: Some dictators in the standard game make "super-fair" transfers, i.e., transfers of more than one-half (e.g., 13%, on average, across various studies in Engel, 2011, 6% in the Standard treatment in Konow, 2010). This is a minority of dictators that is smaller than the fraction of those who make null transfers.

THEOREM 3.1.2: In the standard dictator game, a minority of dictators makes "super-fair" transfers, which are not optimal in the absence of altruism.

PROOF: The assumption that $0 < \mathrm{A}(\overline{\alpha}) - \mathrm{A}(\alpha^*) < 0.5$, where $\alpha^* = \{\alpha | u'(X - \eta) = \sigma\alpha g'(\alpha\eta)\}$ in the standard dictator game, implies there is a minority of dictators, whose optimal transfers are super-fair, since for them $u'(X - \eta) < \sigma\phi f'(0) + \sigma\alpha g'(\alpha\eta)$ when $x = \eta$. Such transfers are never optimal in the absence of the altruism term since, in that case, $u'(X - x) > \sigma\phi f'(0) = 0$.

Note that the part of SF 3.1.2 regarding the fraction of super-fair dictators being smaller than those who make null transfers does not follow from previous assumptions, but this stylized fact is used in the proof of the next claim.

SF/THEOREM 3.1.3: If null transfers are more numerous than super-fair transfers, then the mean transfer in the standard dictator game is strictly between zero and one-half of the stakes (e.g., Camerer, 2003, Engel, 2011).

PROOF: See Appendix 1.

On the basis of this theorem, and the facts, one can disregard super-fair dictators, whenever the focus is on the mean behavior of dictators. In fact, except where otherwise stated, the analysis in this paper refers to mean behavior.

SF/THEOREM 3.1.4: In the standard dictator game, some dictators transfer amounts that equalize or come close to equalizing allocations (e.g., Camerer, 2003, Engel, 2011).

PROOF: According to Assumption 3, the entitlement can be inferred from spectator allocations in the same context. In the standard dictator game, which lacks information about effort, need or other distributive norms, the entitlement reduces to equal splits according to spectator allocations, e.g., Croson and Konow (2009, RZ treatment) and Konow (2000, benevolent/exogenous treatment). Combined with Theorem 2.2.3, transfers closer to equality in these games are consistent with dictators, who have higher values of ϕ.

Note that strict equality does not emerge from fairness preferences alone, since $\phi < \infty$, but it can with the added effect of altruism. An argument for the mass at equality will be discussed later.

Another design, which I call the "tax experiment," consists of dictator games in which a fixed total endowment ($\bar{M}$) is distributed differently across treatments between dictator (X) and recipient (Y), where $X > Y$ and $\bar{M} = X + Y$.

SF/THEOREM 3.1.5: Crowding out is partial (or incomplete) in the tax experiment. Incomplete crowding out means that the average dictator transfer, x, decreases by less than any increase in the recipient's endowment (e.g., Bolton and Katok, 1998, Korenok, Millner and Razzolini, 2017, Cox, List, Price, Sadiraj, and Samek, 2019).

PROOF: See Appendix 1 for the proof of incomplete crowding out, i.e., $-1 < {}^{dx}/_{dY} < 0$.

Fairness alone predicts complete crowding out, i.e., ${}^{dx}/_{dY} = -1$, so this underscores the importance of including altruism in the agent's moral preferences.

Another piece of corroborating evidence for altruism can be found in the study of Crumpler and Grossman (2008). In what I will call the "futile dictator" experiment, the experimenter makes a preset charitable donation, and dictators can also contribute to the charity, but then the experimenter's donation is reduced by the same amount as the dictator's gift, so that the amount received by the charity remains the same. Nevertheless, most dictators (57%) contribute a significant fraction of their endowment (20%, on average). This result is also consistent with our theory, as proven in the following theorem.

THEOREM 3.1.6: Some dictators contribute a positive amount in the futile dictator experiment.

PROOF: See Appendix 1.

Such transfers cannot be explained by fairness but are consistent with agents, whose altruism is

sufficiently strong. The estimates from Crumpler and Grossman not only provide further evidence supportive of altruism (or warm glow) but are also consistent with our assumption that the lower bound on the fraction of agents with altruistic preferences ($\alpha > 0$) is greater than one-half.

In what I call the "subsidy experiment," the dictator's endowment is held constant ($\bar{X}$) while the recipient's endowment is varied across treatments.

SF/THEOREM 3.1.7: Crowding out is partial in the subsidy experiment. Thus, the average dictator transfer, *x*, decreases by less than any increase in the recipient's endowment (e.g., Konow, 2010, Korenok, Millner and Razzolini, 2012).

PROOF: See Appendix 1 for the proof of partial crowding out, i.e., $-1 < {}^{dx}/_{dY} < 0$.

Altruism alone predicts complete crowding out, i.e., ${}^{dx}/_{dY} = 0$, so this validates the inclusion of fairness in the agent's moral preferences.

SF/THEOREM 3.1.8: When information relevant to moral norms is added, stakeholder allocations are significantly positively related to spectator norms involving inequality, including equity/proportionality (e.g., Cherry, Frykblom, and Shogren, 2002, Konow 2000, Konow, Saijo and Akai, 2020, Korenok, Millner and Razzolini, 2017, Oxoby and Spraggon, 2008), need (e.g., Benz and Meier, 2008, Eckel and Grossman, 1996, Konow, 2010, 2019, Muller and Renes, 2017) and efficiency (Almås, Cappelen, and Tungodden, 2020, Charness and Rabin, 2002, Engelmann and Strobel, 2004, Faravelli 2007).

PROOF: This follows from Assumption 3 and Theorem 2.2.5.

This theory can be applied to explain or predict numerous types of stylized facts, including about mean behavior of agents or subgroups of agents, and the distribution of behavior based on salience, fairness types, altruism types, and action sets. But I am postponing discussion of one type of stylized fact until section 7: preference-based masses that are often observed in certain social preference experiments, e.g., a spike at equal splits in many standard dictator games. Preference-based masses are distinct from those due to a constrained action space, which can be explained within the current theory (e.g., SF 3.1.1 above). Later I offer salience-based explanations for preference-based masses by employing a type of moral salience, called point salience, that is different from the set salience otherwise used in this paper.

3.2 Moral Proximity

Effective altruism is a philosophical and social movement that, simply put, advocates for directing charitable resources to where they will do the most good. As compelling as that argument seems, though, it is inconsistent with much charitable behavior: many donors in developed countries prefer local or domestic causes over charities in developing countries, although their donations could make a much bigger difference in the lives of more people with the latter. In this section, I argue that one of the most important and common examples of an anomaly that can be explained by moral salience is what I will call moral proximity. This provides an explanation for the effects on moral behavior of, inter alia, physical distance, homophily, familial relations, friendship, in-groups, or information about the agent or patient. These effects often seem so intuitive that they hardly strike us as anomalous, although it is still sometimes surprising how easily they can be triggered. And yet they are not predicted by most social preference theories, and I am unaware of theoretical accounts that are able to cast all the different examples in a unified framework.

One of the most important practical moral questions is how to identify one's *moral group*, that is, the set of persons to which one is obliged to be moral. Most of philosophical ethics concerns moral principles in general terms and scarcely addresses the question of moral groups (although there are exceptions, e.g., Walzer, 1983). Although equal moral consideration of all seems noble, it is self-evident that moral obligations cannot extend indefinitely, and the boundaries are very much in dispute: some people draw the line at one's family, clan, religious group or ethnic group, some claim we are obliged to our fellow citizens, some to the unborn, some to all people in the world, and some also to animals (which raises the further question of "which animals?"). Even a broad conception of moral group cannot plausibly maintain that all members of that group are equal: surely, the obligation to one's child differs from that to a securities trader in a distant country. The identification of the moral group is paramount to economic policy. For example, suppose one seeks to promote fair earnings (or efficient earnings or any other normative goal). The first order of business is to identify the set of persons whose earners should be targeted: those within a firm, city, county, state, country, the world? There is also the sticky question of which generations to include, which is critical for so many policies including climate change, i.e., do we include only the current generation or also future ones, and, if so, which? There are practical reasons for favoring one answer or the other, but the moral

question must still be factored in, and its resolution is less than obvious.

I will not attempt to resolve these normative questions here, but they are offered as motivation for the importance of the topic and as inspiration for the current descriptive undertaking. The descriptive importance of the topic is suggested by various phenomena, including by the effects of co-workers on productivity, e.g., see Bandiera, Barankay, and Rasul (2010). The focus here is on analyzing of how moral behavior is affected by contextual factors that make the patient's membership in the agent's moral group more or less salient. That is what is meant by moral proximity, and we proceed, as usual when operationalizing moral salience, from a high salience reference point to help identify the properties that affect salience. Assumption 2 in section 2.1 outlined high salience. To elaborate the proximity aspect of that assumption, consider the following assumption.

ASSUMPTION 4: Patients are most morally proximate and, therefore, salient, when, ceteris paribus, they are physically near, personal information about them is abundant and/or stresses their membership in the agent's moral group, they are associated with other proximate persons, others possess abundant personal information about the agent, agent and patient communicate with one another, and agent and patient share traits in common, even ones that might seem superficial and morally irrelevant. In the case of cardinal measures of distance, let the patient, who is most proximate to the agent, have positive measure, $p > 0$ and the additional distance to a more distant patient be non-moral context, n.

The variables listed above are not presented as exhaustive, since the question of what affects perceptions of moral groups is an empirical one. But take first the case of moral proximity and physical distance. Specifically, consider an agent, who may transfer something of material value to a patient, whereby the distance to different patients varies. The factor that varies in this context, C_i, is physical distance, and $m(C_i)$ is a measure of it. Then, define p and n for this cardinal measure as in Assumption 4. Further, denote the total distance $\delta = p + n$ and normalize the measure of the distance of the most proximate patient, i.e., $p = 1$. Then, remembering our specification for moral salience, this can be expressed

$$\sigma = (1 - \bar{\sigma}) \cdot \frac{p}{p+n} + \bar{\sigma} = (1 - \bar{\sigma}) \cdot \frac{1}{\delta} + \bar{\sigma}, \tag{6}$$

which, if $\bar{\sigma} = 0$, reduces to $\sigma = \frac{1}{\delta}$. It is interesting to note that $\frac{1}{\delta}$ also captures a feature of visual salience, viz., the relationship between distance and perceived size: an object at twice the

distance appears half as large. Now we turn to some recent evidence on giving and physical distance.

SF/THEOREM 3.2.1: Agent contributions decrease at a decreasing rate with physical distance to patients (Touré-Tillery and Fishbach, 2017, Dejean, 2020, Kühl and Szech, 2017).

PROOF: This follows from Assumption 4 and Theorem 2.2.2 under the assumptions stated there.

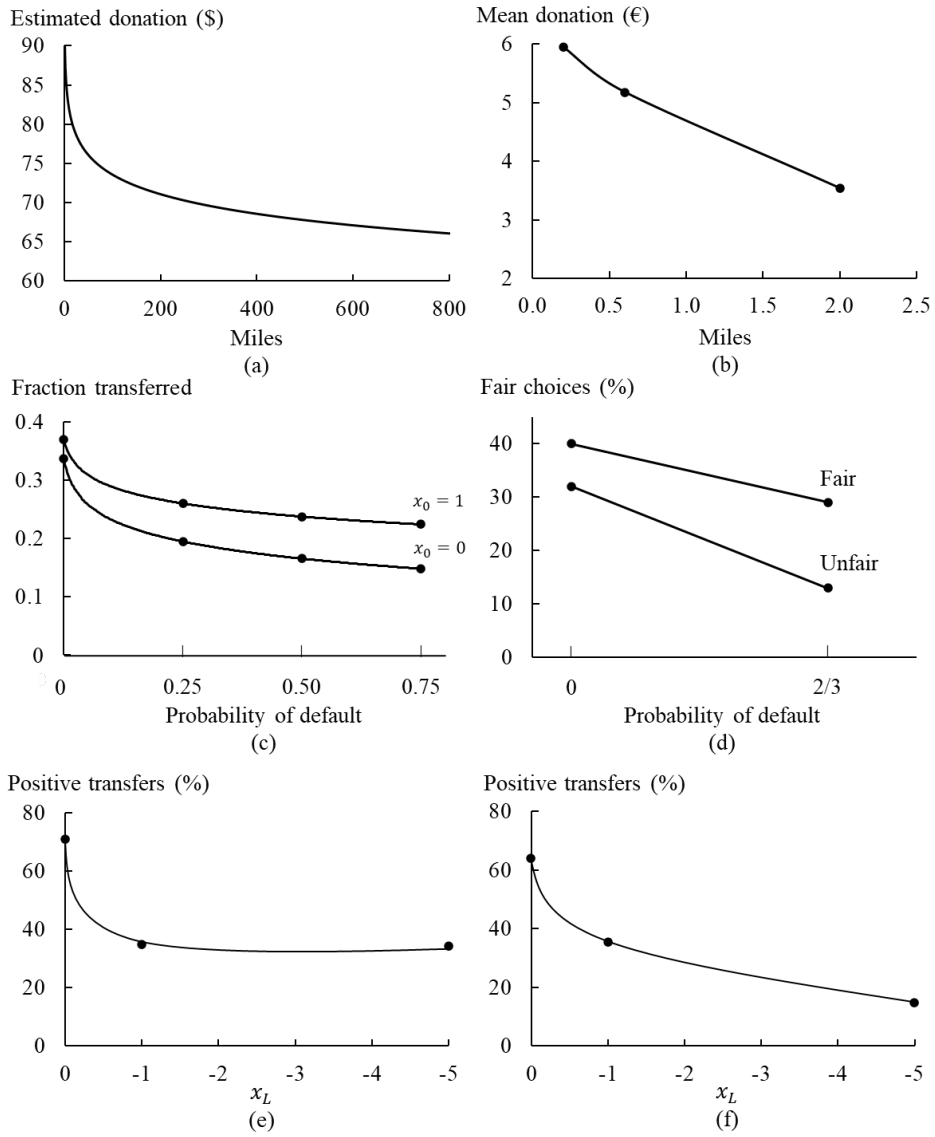

FIGURE 2. – Generosity and Non-moral Context.

Sources: (a) Touré-Tillery and Fishbach (2017) Study 2, (b) Kühl and Szech (2017) laboratory experiment, (c) Andreoni and Bernheim (2009), (d) Grossman (2015) P&O condition, (e) List (2007) baseline, Take $1 and Take $5, and (f) Zhang and Ortmann (2013) baseline, $1 and $5 giving decisions.

Some recent studies find that physical distance influences contributions to others. Touré-Tillery and Fishbach (2017) report that alumni giving to a large private US university is

inversely related to physical distance (Study 2). Figure 2 summarizes the relationships between generosity and non-moral context for six studies. I will return to panels (c) to (f) later in the paper, but note now the consistency of the results with Theorem 2.2.2 across diverse measures of generosity and diverse measures of non-moral context: they are all inversely related, and the generosity measures appear convex in non-moral context, n, in every case where it is possible to detect it, that is, wherever there are more than two levels of n (i.e., except for d). Panel (a) of this figure illustrates the results of a regression based on the data of Touré-Tillery and Fishbach that employs the natural log of distance, which provides a better fit than a linear specification or a non-linear one that adds the square of distance. These results are consistent with moral proximity: donations decrease with physical distance at a decreasing rate. The authors report that the inverse relationship is robust to various controls, including age, income, graduation year, etc.

That said, observational studies cannot rule out omitted variable bias, although they can sometimes shed light on it. Dejean (2020) studies the relationship between rewards-based crowdfunding and physical distance. Using a log specification, he finds investments to decrease with distance at a decreasing rate, viz., they are half as large at twice the distance. Nevertheless, the effect of distance is significantly reduced when social networks are taken into account. Social networks are consistent with a different kind of moral proximity, but this effect weakens claims about physical distance, per se.[8] Such issues are not a concern with the experimental studies of anonymous giving by Kühl and Szech (2017), which permit stronger causal inferences about the effect of physical distance. Their field experiment finds donations to local refugees decrease significantly with distance to their camp, holding other factors constant. Their laboratory experiment comes to similar conclusions about donations to charities, although the relationship is insignificant at longer distances.[9] Figure 2b shows the decrease in average contributions with

[8] One should be cautious, though, about trying to transfer lessons from Dejean's study to the topic of generosity. Rewards-based crowdfunding arguably relates partially to generosity, given that the rewards are often uncertain and not commensurate with the investments, but there is an expectation of some "reward," even if only a thank you note, which leaves a role for the confounding force of strategic self-interest. In addition, the dependent variable is the number of contributions rather than their value.

[9] The effect at longer distances of up to 6000 miles is likely confounded by other factors. Participants report lower feelings of responsibility toward more distant recipients, consistent with moral proximity, but their contributions are not significantly related to distance. The authors attribute this to participants failing truly to have constant beliefs across distances despite the authors attempts to hold all else constant, e.g., through claims about similar per capita GDP. That seems plausible, since Germany, where the study was conducted, has one of the highest per capita GDPs in the world, and their questionnaire results show a much higher focus on people at longer distances being in need. In addition, Germany has one of the lowest levels of inequality in the world, so even if subjects believe that distant

distance for their laboratory experiment. Although the convexity appears subtle in this case, that is due to the short distances, and the change in the slope is actually comparable to that in Dejean and greater than that in Touré-Tillery and Fishbach for comparable distances.

Many of the other factors that influence moral proximity have often been characterized as "social distance" (in the pre-COVID-19 sense of the term). For example, transfers rise, when even limited information about the dictator is provided to the experimenter (Hoffman, McCabe and Smith, 1996), the recipient (Bohnet and Frey, 1999, Grossman, 2015), or both (Alevy, Jeffries, and Lu, 2014). In fact, giving rises, even if the mere existence of a dictator, who remains anonymous, is revealed to the recipient (Dana, Cain and Dawes, 2006). Even three dots on a screen in a "watching eyes" position, instead of a neutral position, can increase dictator transfers (Rigdon et al., 2009). Some of these effects can plausibly be attributed, at least in part, to social image concerns, even under anonymity (see the discussion in the following section of Andreoni and Bernheim, 2009). Nevertheless, social image does not easily explain other factors that fall under the rubric of moral proximity. Dictator giving rises, if the recipient reveals one-way his/her identity to the dictator (Bohnet and Frey, 1999), indeed, even if only the recipient's family name is revealed (Charness and Gneezy, 2008). Conversely, dictator gifts fall, if it is revealed that the recipient is a member of an out-group (Whitt and Wilson, 2007), and Candelo, Eckel and Johnson (2018) report that dictator transfers to a family member are greater than those to a community group or stranger. In addition, transfers increase, if recipients can send a message to the dictator (Bohnet and Frey, 1999, Ellingsen and Johannesson, 2008, Xiao and Houser, 2009).

Unlike physical distance, these social distance factors typically do not lend themselves to cardinal measures of p and n. But they provide plausible examples of moral proximity, and we will occasionally return to this effect in later sections. Moreover, the direction of the effect of a factor on salience in these cases is obvious, and that suffices where we employ such qualitative variables to explain categorical changes due to salience.

3.3 Moral Uncertainty

Uncertainty is present in virtually all economic decisions, and it can often be managed to some degree. Nevertheless, people sometimes use uncertainty as an excuse to avoid costly

recipients really do enjoy the same average income, levels of inequality elsewhere are surely higher, meaning that the distant poor are likely needier than the local poor.

actions that are otherwise justified on both economic and moral grounds, such as taking steps to address climate change (Finus and Pintassilgo, 2013). Many studies have demonstrated the relevance of uncertainty to economic decision making, including in economics experiments involving moral preferences, e.g., Bolton, Brandts and Ockenfels (2005), Brock, Lange and Ozbay (2013), Cappelen, Konow, Sørensen and Tungodden (2013), Rey-Biel, Sheremeta and Uler (2018), Van Koten, Ortmann and Babicky (2013), and Zizzo (2003). In particular, the controlled methods of experiments can help show that the reduction in moral conduct with increased uncertainty is associated with moral preferences themselves and cannot be dismissed as being due solely to other forces, such as risk preferences.

Moral uncertainty refers to the depressing effect on moral salience because of uncertainty in the context about the agent's choice. Assumption 5 specifies the kind of experimental uncertainty considered here as well as its relationship to moral salience.

ASSUMPTION 5: In the "uncertainty game" allocations may be randomly determined by an agent or by default, the latter because either the agent is randomly precluded from choosing allocations or the agent's choice is not randomly chosen for realization. The probability of such a default constitutes non-moral context, n, where $n \in [0,1]$. Moral context has some positive measure, $p > 0$, the value of which depends inversely on the sensitivity of σ to n in the selected context. Baseline moral salience, $\bar{\sigma}$, depends inversely on the unfairness of the default.

Thus, this assumption means that the possibility that allocations will not be based on the agent's choice reduces the moral salience of that choice. Moreover, salience is further reduced as the probability increases that the agent's choice will not matter and as the default becomes less fair. In this framework, the moral measure, p, represents a certain implicit moral context and a parameter that effectively calibrates the sensitivity of σ to n. Similarly, the assumption about $\bar{\sigma}$ captures the concept that a less fair (fairer) default lowers (raises) moral salience given that $\sigma = (1 - \bar{\sigma}) \cdot \frac{p}{p+n} + \bar{\sigma}$. Assuming a specific form for the utility function, p and $\bar{\sigma}$ might be estimated empirically, but the theoretical analysis here does not depend on any particular values for these parameters beyond Assumption 5.

Although numerous economics experiments have investigated uncertainty, I am aware of only a small number with designs suitable to the criteria considered here. As usual, the design must involve non-strategic decisions, and probabilities should be manipulated at two levels at a minimum. Some studies that satisfy these conditions must nevertheless be ruled out because their

design activates risk preferences (e.g., Krawczyk and Le Lec, 2010) or fairness preferences over risk because subjects choose levels of risk-taking (e.g., Cappelen et al., 2013). I focus on two studies that satisfy all requirements while representing two different and important categories of moral uncertainty. They lead to the following stylized fact and theorem.

SF/THEOREM 3.3.1: Dictator transfers decrease at a decreasing rate with the probability of the default. The fairer the default, the greater the mean transfer and the greater (less negative) $\partial x/\partial n$ (Andreoni and Bernheim, 2009, Grossman, 2015, P&O treatment).

PROOF: The claims about transfers follow from Assumption 5 about *n* and Theorem 2.2.2 under the assumptions stated there. The claims about the effects of the fairness of the default follow from Assumption 5 about $\bar{\sigma}$, Theorem 2.2.1, Theorem 2.2.2, and the facts that $\frac{\partial \sigma}{\partial \bar{\sigma}} = \frac{n}{p+n} > 0$ and $\frac{\partial^2 \sigma}{\partial n \partial \bar{\sigma}} = \frac{p}{(p+n)^2} > 0$.

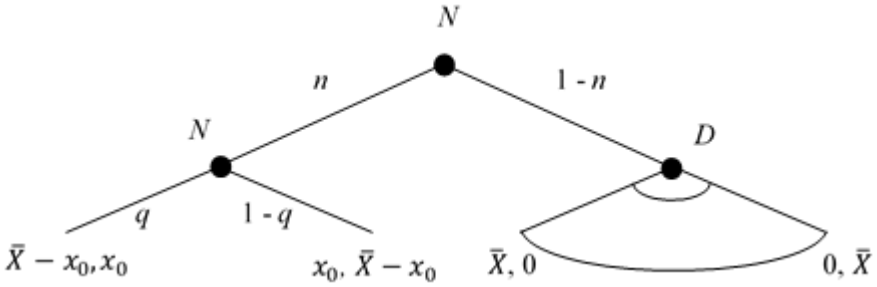

FIGURE 3. – Uncertainty game of Andreoni and Bernheim (2009).

Andreoni and Bernheim (2009) report the results of a dictator game, the design of which is illustrated in extensive form in Figure 3. Nature (N) first decides whether the Dictator (D) or Nature will allocate stakes, $\bar{X}$, of $20 between D and a recipient (R). Nature allocates with probability *n*, whereby this probability varies across four levels within subjects, viz., $n \in \{0,0.25,0.5,0.75\}$. If Nature allocates, there is an equal chance, $q = 0.5$, either that $\bar{X} - x_0$ goes to D and x_0 to R or that x_0 goes to D and $\bar{X} - x_0$ to R, where $x_0 \in \{0,1\}$ is varied between subjects. With probability $1 - n$, the allocations will follow the decision of D, who can choose any amount, $x \in X = [0,20]$. Consider panel c of Figure 2, which was introduced in the prior section and shows the results of regressions of the fraction of stakes transferred by D to R on *n* for $x_0 = 0$ and $x_0 = 1$, separately. The slope and rate of change in fractional transfers are consistent with the predictions: transfers decrease with the probability of the default at a decreasing rate. The height and slope of transfers are also consistent with the predicted effects of

the fairness of the default: the curve for the fairer $x_0 = 1$ treatment displays higher transfers and a flatter slope than the one for $x_0 = 0$.

Note that, if D finds him/herself in this branch of the game tree, the decision is entirely ex post. That is, the uncertainty has been resolved, and D knows with certainty that his/her transfer will be realized. This type of uncertainty should not matter in standard social preference theories. Andreoni and Bernheim make a persuasive case, however, for their theory of social image that is consistent with the results of this experiment. My aim here is not to fault the social image argument, which I find credible, but rather to provide an explanation based on moral salience that predicts these and additional results, while noting that the two accounts are not mutual exclusive.

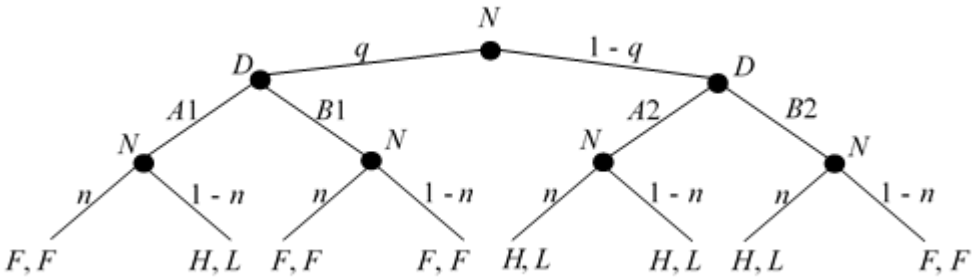

FIGURE 4. – Uncertainty game of Grossman (2015).

The moral uncertainty argument does not require that the agent's action set, X, be subject to uncertainty, because moral salience can be affected by information about uncertainty in the decision context, Y. This explains how non-moral elements of Y can reduce moral salience, even when, as in this experiment, decisions are ex post. According to this account, an experiment in which one subject might have been randomly chosen to receive an unfair share of (almost) all of the stakes diminishes the prominence of moral considerations and, therefore, of moral preferences, more so, as the probability of the default increases and as the default becomes less fair. In addition, the moral salience explanation has five uniquely attractive features. First, it can explain not only the decrease in x with n but also the decreasing rate of change. Second, it predicts the decreased rate of change (i.e., flatter slope) at higher levels of x_0. Third, it explains the increase in x with x_0 among those Ds giving more than x_0 due to higher fixed moral salience. This effect is not predicted by, and, in fact, is inconsistent with, social image, which predicts that the higher transfer, when $x_0 = 1$, should be due solely to the shift by Ds, who would otherwise give zero at $x_0 = 0$. Fourth, the theory is simple and parsimonious. Fifth, moral salience is consistent with a wide range of other anomalies that are not predicted by alternative accounts

such as social image.[10]

Grossman (2015) reports a variation on a dictator experiment designed to test his theory of social-image and self-image. Dictators not only make ex post decisions, which follow the resolution of some uncertainty as in Andreoni and Bernheim, but also face ex ante uncertainty. The experimental design in illustrated is Figure 4. This is a binary dictator game with only two possible pairs of payoffs to D,R of (H,L) or (F,F), where $0 \leq L < F < H$ and $L + H < 2F$, specifically, in Grossman (2015), $H = 7$, $F = 5$, $L = 1$. Nature first determines with equal probability, $q = 0.5$, which of two games the dictator will play, 1 or 2, and this random assignment is common knowledge. These games differ based on whether the default is Fair, i.e., (F,F), in game 1 or Unfair, i.e., (H,L), in game 2. Subjects are randomly assigned one of two probabilities of the default obtaining, $n \in \{0, \frac{2}{3}\}$, which is also common knowledge. Dictators choose the payoffs ex ante in the event their decision is chosen, either the Unfair option A (A1 in game 1 or A2 in game 2) or the Fair option B (B1 in game 1 and B2 in game 2).

Ds make these decisions in three treatments of a between-subjects design that differ with respect to the information that is available to Rs. They can observe the D's choice (C), the outcome (O) or both the probability and outcome (P&O). Comparing variation in the fraction of Fair choices with probabilities in the three treatments, the results produce little support for self-image and mixed results on social image. On the other hand, the results are significantly consistent with most predictions of moral uncertainty in two of three treatments (C, P&O) and insignificantly opposite it in the third (O). This last fact is perhaps because social image concerns muddy the waters somewhat, especially, where they are predicted to do so in O. The results of the treatment with the information conditions closest to the standard dictator game (viz., P&O) are summarized in panel d of Figure 2. This shows that the average transfer to R, or equivalently here, the fraction of Ds choosing Fair, decreases with *n*, and the fairer default results in higher average transfers and a flatter slope. Note that there is one claim of Theorem 2.2.2 to which the Grossman study cannot speak: since these experiments only vary the probability of the default at two levels, these results cannot shed light on the rate of change of *x* with *n*.

[10] The later section on moral point salience presents an additional salience-based argument for the higher average transfers when $x_0 = 1$ than when $x_0 = 0$ as well as for masses at those values.

4. Applications of Virtue Preferences

This section applies the model that includes virtue preferences to classic results about reciprocity and to anomalous findings about the so-called outcome bias.

4.1. Reciprocity

Reciprocity refers to a type of behavior, where people return kindness with kindness, called positive reciprocity, or unkindness with unkindness, called negative reciprocity. Such behavior has been found in numerous experimental designs, including with the seminal gift exchange game of Fehr, Kirchsteiger and Riedl (1993), the trust game of Berg, Dickhaut and McCabe (1995), the triadic design of Cox (2004), and the moonlighting game of Abbink, Irlenbusch and Renner (2000). These results can be summarized in the following statement.

SF 4.1.1: Stakeholders sanction (Güth, Schmittberger and Schwarze, 1982, Berg, Dickhaut and McCabe, 1995, Cox, 2004, Abbink, Irlenbusch and Renner, 2000). Moreover, they sanction asymmetrically, punishing low generosity more strongly than they reward high generosity (Croson and Konow, 2009, Cushman, Dreber, Wang and Costa, 2009, Offerman, 2002).

Further studies show that subjects exhibit generalized reciprocity, acting not only when they are the objects of kindness or unkindness but also as third parties sanctioning kindness or unkindness by others toward others, e.g., Almenberg, Dreber, Apicella and Rand (2011), and Fehr and Fischbacher (2004).

A critical question is the extent to which such behavior reflects reciprocal altruism, i.e., a preference to reward or punish, or some other motive (see Sobel, 2005, for an excellent theoretical treatment of types of reciprocity). As an illustration of this distinction, consider the ultimatum game, in which a "proposer" proposes a division of a fixed sum of money between him/herself and a "responder," and the responder either accepts, and the sum is divided as proposed, or rejects, in which case both earn nothing (Güth, Schmittberger and Schwarze, 1982). The subgame perfect Nash equilibrium under that standard assumptions of rational, self-interested agents is for the responder to accept any amount and for the proposer, therefore, to offer the minimum amount. Nevertheless, the results of hundreds of replications show that proposers typically offer non-negligible positive amounts and responders often reject offers of less than one-half (Camerer, 2003). But there are various alternative motives at play in this game.

For example, responders might reject for purely distributive, or fairness, reasons and not to punish the proposer, or the responder might care about efficiency, or the responder might have altruistic preferences toward the proposer. Even if the responder wishes to punish the proposer, though, even a comparatively generous proposer's intentions are clouded by motives other than fairness, such as a self-interested desire to avoid rejection, which is further confounded by risk preferences. Indeed, comparison with other games imply decisions in the ultimatum game result from a confluence of motives, e.g., Forsythe , Horowitz, Savin and Sefton (1994).

For those reasons, I discuss a non-strategic dictator experiment on reciprocal motives. Croson and Konow (2009) introduced a two-stage dictator game in which a dictator first chooses one of six divisions with a recipient, {(10-0), (8-2), (6-4), (4-6), (2-8), (0-10)}, of a sum of money, $X = 10$. Then, there follows a previously unannounced second stage, in which a different subject chooses a division between the same subjects of an additional sum of money, $Z = 20$, in any integer amounts. In one experimental condition, the second stage dictator and recipient are the first stage recipient and dictator, respectively. For clarity, I will refer to them according to their first stage roles as D and R, respectively. In this condition, R is a *stakeholder*, or party to the allocations. In another condition, the second stage allocator is a third party, or *spectator*, who is paid a fixed sum to allocate *Z* between D and R. Another pair of treatments is identical to these two with stakeholder and spectator versions, except the first stage allocation is not chosen by anyone but rather is randomly assigned, so this is a 2×2 between-subjects design. The strategy method is employed for second stage allocations: all second stage allocators choose a division of *Z* for each of the six possible first stage divisions. The variable of interest is the allocation decisions about *Z* by the second stage allocators based on whether they were themselves a stakeholder or spectator and on whether the first stage division was chosen by a dictator or randomly determined.

Take first the case of the stakeholder, R, who is endowed with the amount received from the first stage, *x*, plus the second stage sum, *Z*, and can transfer any amount, $z \in [0, Z]$, to D (note *f* refers here to D as the second stage recipient). This R's utility function can be written

$$U = u(x + Z - z) + \sigma \cdot f\big(\phi(z - \eta_z - \theta \cdot r(\hat{\gamma} - \tilde{\gamma}) \cdot x')\big) + \sigma \cdot g(\alpha z)$$

where the entitlement of the current recipient is $\eta_z = {}^{Z}/_{2} - X + x + \eta_x$. That is, D is entitled to one-half of the current endowment, ${}^{Z}/_{2}$, since this is a simple D game. In addition, η_z reverses

any inequity in how much D took in the first stage, $X - x - \eta_x$, where $\eta_x = {}^{X}/_{2}$ is the fair division of the first stage stakes, again equal splits because this is a simple D game. We can also specify virtue preferences more precisely for this experiment. As explained in section 2.3, threshold generosity, $\tilde{\gamma}$, is the break-even level of generosity for reward or punishment and depends on moral salience in the reference state, $\tilde{\sigma}$. But since the choice of reference state is arbitrary and salience depends on many aspects of moral and non-moral context, we can define the reference state to correspond to one in which moral salience is at the same level as in the first stage of the dictator game at hand. In that case, threshold generosity, $\tilde{\gamma}$, equals the threshold transfer of D in the first stage, $\tilde{x}$. Similarly, R's estimate of D's generosity, $\hat{\gamma}$, then corresponds to D's actual generosity in the first stage, x.[11] Finally, the scale is the fair division of the second stage sum, i.e., $x' = {}^{Z}/_{2}$. Then, R's utility function in the second stage can be written

$$U = u(x + Z - z) + \sigma \cdot f\left(\phi\left(z - {}^{Z}/_{2} + {}^{X}/_{2} - x - \theta \cdot r(x - \tilde{x}) \cdot {}^{Z}/_{2}\right)\right) + \sigma \cdot g(\alpha z).$$

Now we come to the following theorem about stakeholders in this experiment.

THEOREM 4.1.1: In the two-stage dictator game, second stage allocators, who are stakeholders, partially adjust for first stage transfers that are random, i.e., $0 < {}^{dz}/_{dx} < 1$, but some stakeholders sanction first stage dictators, i.e., ${}^{dz}/_{dx}$ increases, when those dictators choose first stage allocations.

PROOF: See Appendix 1.

When first stage endowments are random, the sanction term, $\tilde{z}$, drops out, and the partial adjustment of z to x is merely an application of Theorem 3.1.5 to the two-stage dictator game. When first stage dictators reveal their character through their choices, however, this term increases to include sanctioning.

Note that one part of SF 4.1.1 that is not claimed in Theorem 4.1.1 is asymmetric sanctioning. Stakeholder allocations do not produce a clear measure of this asymmetry, given the confluence of additional motives, including material self-interest, inequity aversion and altruism, so we turn now to spectators, whose decisions are predicted to generate an undistorted measure. The utility function of a spectator in this experiment can be written

[11] To be exact, $\hat{\gamma} = x$ for interior solutions, but further specification of $\hat{\gamma}$ is needed in the case of corner solutions. This refinement is unnecessary for the current focus on mean behavior, but it will be addressed in section 5.2, where it provides insight into an additional finding that becomes apparent in the design discussed there.

$$U = u(\bar{z}) + \sigma f\left(\phi\left(z - {}^{Z}/_{2} + {}^{X}/_{2} - x - \theta \cdot r(x - \tilde{x}) \cdot {}^{Z}/_{2}\right)\right).$$

where $\bar{z}$ represents the fixed payment the spectator receives for making this decision. Remember that no altruism term is included in a spectator's utility function, since the relationship is impartial rather than personal.[12] Theorem 4.1.2 follows.

THEOREM 4.1.2: In the two-stage dictator game with randomly assigned first stage allocations, second stage allocators, who are spectators, equalize, adjusting completely for first stage transfers that are random, i.e., ${dz}/_{dx} = 1$. When dictators choose first stage allocations, some spectators sanction them, i.e., ${dz}/_{dx} > 1$, and some equalize. Those who sanction have different thresholds, $\tilde{x}$, and sanction, on average, asymmetrically, punishing more strongly than they reward.

PROOF: See Appendix 1.

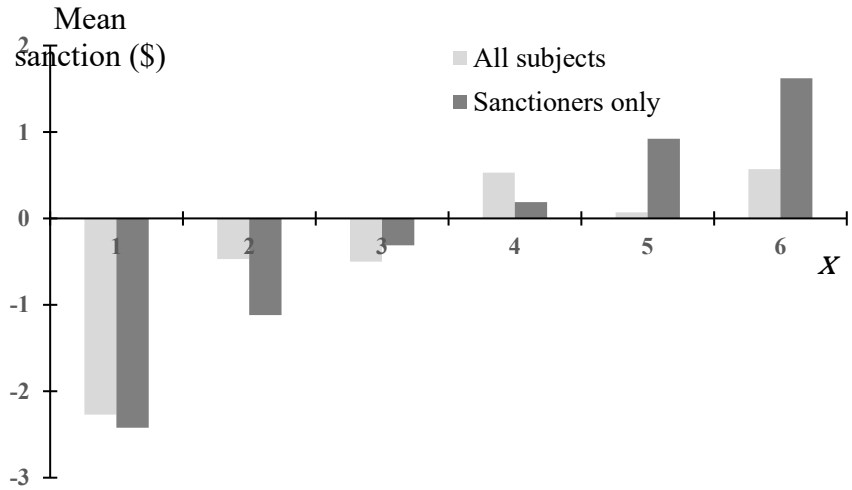

FIGURE 5. – Sanctioning in Croson and Konow (2009).

Spectator allocations are motivated solely by fairness and virtue preferences. Following SF 3.1.4, fairness reduces to equality in a simple dictator game of this sort. When first stage allocations are random, spectators adjust *z* to *x* one-for-one, but, when they are chosen by first stage dictators, virtue preferences kick in, and spectators sanction. The asymmetry follows from the fact that ${d^2z}/_{dx^2} < 0$ due to the concavity of *r*. Figure 5 illustrates results from Croson and

[12] Note also that the fairness preference is formulated with respect to the first stage dictator, because that is the subject, who has revealed something about his character. A term could be added for fairness toward the first stage recipient, but unfairness toward one first stage subject is simply mirrored by unfairness in the opposite direction toward another, and the conclusions are qualitatively unaffected, so I avoid this clutter as in previous analysis of spectator allocations in dictator games, e.g., Konow (2000).

Konow of spectator reward or punishment as the mean amounts by which their transfers to the first stage dictator, *z*, fall short of or exceed equalizing transfers for each level of transfer chosen by first stage dictators. The mean sanctions of all spectators are illustrated in the light columns, but only 50% of spectators sanction, consistent with the assumption that some agents place zero weight on sanctioning ($\Theta(0) > 0$). Another 37% equalize, and the other 13% cannot be put into either category.[13] The mean reward and punishment of only those who sanction are illustrated in the darker columns. Figure 5 suggests an asymmetry, which is corroborated in more formal analysis in Croson and Konow and is generally consistent with third party sanctioning in other studies, such as Almenberg et al. (2011) and Fehr and Fischbacher (2004). Moreover, one observes the assumed heterogeneity in thresholds: 27% of spectators stop punishing (and thereafter either equalize or begin rewarding) at a first stage transfer of 0, 40% at 2, 20% at 4, and 13% at 6.

4.2. Outcome Bias

Numerous studies across multiple disciplines have established a preference for rewarding or punishing individuals based on uncontrollable (or brute) luck, including in politics (e.g., Healy, Malhotra, and Mo, 2010), sports (e.g., Kausel, Ventura and Rodriguez, 2019), and CEO compensation (e.g., Bertrand and Mullainathan, 2001). In the law, someone, who kills another accidentally, can be sentenced more harshly than someone, who meant to kill another but failed (Cushman et al., 2009). In economics, this so-called *outcome bias* has also been extensively studied experimentally, especially, in the context of the principal-agent problem (e.g., Charness and Levine, 2007, Rubin and Sheremeta, 2016). Although the optimal contract rewards effort and disregards luck, people sanction luck, even when decision makers are clearly not responsible (Gurdal, Miller and Rustichini, 2013) and even when those who are sanctioning are third parties (Brownback and Kuhn, 2019). Outcome bias is also enigmatic from the perspective of reciprocity theories in behavioral economics, which conceptualize reward and punishment in terms of intended consequences (e.g., Rabin, 1993, Falk and Fischbacher, 2006). Outcome bias has additionally been a focus of philosophical deliberation (e.g., Williams, 1981).

In this section, I argue that, although outcome bias is inconsistent with optimal contracts

[13] The comments of the 13% in the post-experimental questionnaire indicate that half misunderstood their task and that the other half believed (mistakenly) that some aspect of the decisions was strategic.

and reciprocity theories, it is consistent with moral intuition and virtue preferences. Indeed, in a more general specification of virtue preferences, outcome bias is not a bias, at all. Consider the intuition: two individuals, who are operating a motor vehicle, run a stop sign, the one without consequence and the other causing the death of a family. Do we legally, and should we morally, really hold the two equally accountable? In the one case, the driver might pay a fine of a few hundred dollars, whereas, in the other case, the driver can be found guilty of manslaughter and serve jail time. I claim that "outcome bias" is a feature, not a bug, of moral preferences: the relevant sanctioning motive is with respect to intent coupled with consequential action, as treated in virtue preferences and its philosophical inspiration, virtue ethics. Specifically, the difference in sanctions reflects the intuition exemplified above, which can be incorporated by scaling virtue preferences according to whether or not the intended outcome obtained. Consider the following stylized facts from economics experiments on this topic.

SF 4.2.1: When an agent can choose actions with uncertain outcomes, others sanction the agent, meaning these rewards and punishments are not explained by distributive preferences alone. Sanctions are asymmetric: low generosity is punished more strongly than high generosity of equal magnitude (Cushman et al., 2009, de Oliveira, Smith and Spraggon, 2017).

SF 4.2.2: Sanctions are based not only on the chosen action (and its expected outcome) but also the realized outcome, for which the agent is not responsible. The sanctions for actions leading to outcomes that are both expected and realized are greater than those for outcomes that are expected but not realized (Charness and Levine, 2007, Gurdal, Miller and Rustichini, 2013, Rubin and Sheremeta, 2016), and both stakeholders as well as third parties exhibit this behavior (Brownback and Kuhn, 2019, Gino, Shu and Bazerman, 2010, Sezer, Zhang, Gino and Bazerman, 2016). The sanctions for realized outcomes that agree with expected outcomes are roughly the same as those for the same outcomes chosen with certainty (Cushman et al., 2009).

Despite considerable variation in features of these experiments, including in the role of uncertainty, effort, information and payoff functions, the findings are quite consistent. Given these design differences and the focus of the current analysis, therefore, I will analyze a hybrid design that captures elements of different studies and fits the focus here on non-strategic decisions. In what I will call the "fair luck game," there are two stages, whereby first stage Ds

select an option from a discrete set of risky choices that differ in their expected fairness, and then, in a second stage, the Rs may sanction the Ds based on the latter's choices and the realized payoffs. This game is similar to Cushman et al. (2009), except the risky options number two, as in Sezer et al. (2016), rather than three, both in order to simplify the analysis and because there is empirically little difference between the second and third choices. Specifically, suppose the first stage payoffs to D,R can be either "fair" (F,F) or "unfair" (H,L), where $0 \leq L < F < H$ and $L + H = 2F = X$. The first stage D makes a risky choice, $x \in \{f, u\}$, involving probability $q > 0.5$, which results in expected payoffs to R, EX^f and EX^u, respectively, of

$$EX^f = qF + (1-q)L$$

$$EX^u = (1-q)F + qL$$

where, obviously, $EX^f > EX^u$.

The subject matter of outcome bias is sanctioning, so we focus on the second stage. A potential problem is that first stage dictators anticipate sanctioning in the second stage, in which case their choices might be distorted by strategic self-interest. This can be obviated by a previously unannounced second stage, as in Croson and Konow (2009), or by a negligible probability (viz., 0.1) that the sanction will be implemented, as in Cushman et al. (2009). Formally, the analysis treats all decisions as non-strategic, but the results that follow are robust under more general experimental conditions as long as first stage choices preserve the ranking of Ds by their generosity. At any rate, in the second stage, the first stage R may sanction the first stage D by adding money to, or deducting money from, D's payoff (I will refer to them consistently as R and D according to their roles in the first stage). This game requires specification of moral character, or generosity in this context, γ, to accommodate decisions under uncertainty. Analogous to deterministic decisions, consider a reference state in which the first stage D may choose a benefit to the R but now let γ denote the expected payoff to R, which is distributed on the interval $[\underline{\gamma}, \overline{\gamma}]$. In addition, suppose R expects there is a D type, γ^i, who is indifferent between f and u, where $\underline{\gamma} < EX^u < \gamma^i < EX^f < \overline{\gamma}$. Whereas γ represents notional generosity, effective generosity is constrained in this game to a binary choice between EX^u and EX^f. Thus, R's estimate of D's notional generosity, $\hat{\gamma}$, is either $\hat{\gamma}^f$ or $\hat{\gamma}^u$, depending on D's choice of either f or u, which, respectively, equal

$$\hat{\gamma}^f = \int_{\gamma^i}^{\overline{\gamma}} \gamma\, \rho(\gamma) d\gamma \Big/ \int_{\gamma^i}^{\overline{\gamma}} \rho(\gamma) d\gamma$$

and

$$\hat{\gamma}^u = \int_{\underline{\gamma}}^{\gamma^i} \gamma\, \rho(\gamma) d\gamma \Big/ \int_{\underline{\gamma}}^{\gamma^i} \rho(\gamma) d\gamma.$$

Note that $\hat{\gamma}^u < \hat{\gamma}^f$, and it is further assumed that $\hat{\gamma}^u \leq \tilde{\gamma} \leq \hat{\gamma}^f$, i.e., R's threshold for sanctioning D lies within the interval of D's estimated generosity.

Now consider R's payoff in the second stage. In the fair luck game, R receives a fixed amount from the first stage, x_r, i.e., the realization of D's choice in the first stage. Next is the question of the price R pays to sanction D. In the two-stage D game discussed in section 4.1, R allocates a fixed sum between D and R. That is, letting z be the amount added to or subtracted from D's payoff in the second stage and Y the amount, as a result of z, that is added to or subtracted from R's payoff, then in the two-stage D game, $^{dY}/_{dz} = -1$. The fair luck game is different in that sanctioning is free and produces neither gains nor losses for R, that is, $^{dY}/_{dz} = 0$. Unlike the prior case, then, there are efficiency implications of R's decision. As previously noted, a large volume of research finds evidence that social preferences include, and sometimes are even dominated by, efficiency concerns (e.g., Charness and Rabin, 2002, Engelmann and Strobel, 2004). The results of experiments on sanctioning discussed here are also consistent with the idea that, when agents can sanction and $^{dY}/_{dz} = 0$, efficiency preferences crowd out other allocative preferences (e.g., Bartling et al., 2014, Bartling and Fischbacher, 2012, Cushman et al., 2009).[14] I model this effect with the parameter $\beta = -\,^{dY}/_{dz}$, whereby, in the cases considered here, $\beta \in [0,1]$. The D's entitlement in the second stage is assumed to be

$$\eta_z = \left(\frac{1+(1-\beta)b}{2}\right) Z - \beta(X - x - \eta_x).$$

Thus, in the prior two-stage D game where $\beta = 1$ and efficiency plays no role, $\eta_z = {}^{Z}/_{2} - X + x + \eta_x$, which, as before, splits the second stage sum equally and corrects any inequity from the first stage. In the current fair luck game where $\beta = 0$ and only efficiency matters, $\eta_z = \left(\frac{1+b}{2}\right) Z$, where $0 < b < 1$. The highest possible payoff is Z, which equals zero in games with only punishment, and $\eta_z = 0$. In games with reward, $Z > 0$ and $\frac{1}{2} Z < \eta_z < Z$ because $0 < b < 1$.[15]

[14] In fact, they are so strong in Cushman et al. that 17% of second stage allocators transfer the maximum regardless of first stage actions or outcomes.

[15] The assumption that $b < 1$ is not critical, but it makes the theoretical predictions consistent with the fact that, in

The R's utility function in the fair luck game can, therefore, be written

(7) $U = u(x_r) + \sigma f(\phi(z - \eta_z - \theta \cdot r(\hat{y} - \tilde{y}) \cdot x')) + \beta\sigma g(\alpha z)$.

This reflects R's fixed payment from the first stage. In addition, as stated, when agents sanction and $\beta = 0$, efficiency is assumed to crowd out other allocative preferences, so the effect of β on altruism is technically included but superfluous, in this case. The final step in specifying virtue preferences to accommodate outcome bias involves the scale, x'. So far, choices, if implemented, have had certain consequences in the cases considered, and the scale was defined as the patient's entitlement. Now the consequences of choices are uncertain, and outcome bias is a reflection of the dependence of sanctions not only on choices but also outcomes. In the fair luck game, the fair allocation from the first stage sum of X is $\eta_x = F$, so it makes sense for this to be the scale, when intended and realized outcomes align, whether choices are under certainty or uncertainty. In this case, I will write the choice and realized outcome as the pair $(x, x_r) \in \{(f, F), (u, U)\}$. What about the cases, when intended and realized outcomes do not agree, i.e., $(x, x_r) \in \{(f, U), (u, F)\}$? It is natural, in this case, to think in terms of expected outcomes. As attempted murder is not punished as harshly as murder, so also the scale responds to the difference between expected and realized outcomes. And as attempted murder is punished more severely than attempted robbery, so also the scale responds to differences in expected outcomes. I propose defining the scale in these cases as the expected value from the choice, i.e., EX^f if the choice is fair, and EX^u if the choice is unfair. Then, the scale of sanctions can be defined as

$$x' = \begin{cases} F \; if \; (x, x_r) \in \{(f, F), (u, U)\} \\ EX^f \; if \; (x, x_r) = (f, U) \\ EX^u \; if \; (x, x_r) = (u, F) \end{cases}.$$

These two theorems follow, the proofs of which may be found in the Appendix.

THEOREM 4.2.1: In the fair luck game, agents in the second stage sanction, i.e., they allocate beyond what is called for by distributive preferences alone. Specifically, they reward first stage dictators more, or punish them less, for choosing f than for choosing u, ceteris paribus, i.e., for a given scale, x'.

THEOREM 4.2.2: Choices are sanctioned, even when the intended outcomes do not obtain, but fair choices are rewarded more strongly and unfair choice punished more strongly when

such experiments, some second stage allocators set their goal above equality but, on average, somewhat below the maximum possible reward Z.

realized outcomes are aligned with choices. Sanctions for realized outcomes that agree with expected outcomes are the same as for the same outcomes chosen with certainty.

These theorems predict most of the stylized facts as well as the specific findings of Cushman et al. while adding a theoretical underpinning for them in terms of fairness preferences. The theory is also consistent with the asymmetry in sanctioning in SF 4.2.1 from the concavity of *r*, but this cannot be proven for the fair luck game, given that it produces only a binary signal of preferences. Further corroborative evidence of this specification of virtue preferences will be discussed in section 6.2 on willful ignorance and section 6.3 on delegation.

5. Helping and Harming

This section analyzes anomalies that involve the distinction between helping versus harming others. The specific manifestations of helping considered involve increasing the payoffs of others and of harming decreasing the payoffs of others.

ASSUMPTION 6: In contexts with high moral salience, $p > 0$ and is increasing the set of helping choices, and $n \geq 0$ and is increasing the set of harming choices.

Although fairness preferences are often involved in the findings discussed in this section, the moral preference that relates most directly to helping and harming is altruism, so it is not surprising that explanations for some of the anomalies here rest on differences across agents in altruism. One might think of altruism salience as being more important than fairness salience in this section, although that is not a necessary assumption for any of the claims here.

5.1. The Taking Effect

During civil disturbances and natural disasters, otherwise law-abiding citizens sometimes join in looting (e.g., Green, 2007, Khazan, 6/2/2020, *The Atlantic*, Quarantelli and Dynes, 1968). Scholars have offered many explanations for such behavior, but the results of economics experiments have demonstrated that extrinsic incentives, such as reduced expectations of being punished, cannot, at least solely, explain such abandonment of morals, when opportunities to take from others are offered. Consider an anonymous between-subjects dictator game, in which Rs are also endowed but at a lower level than Ds, and Ds are permitted not only to give in a "Give" treatment but also to take in an otherwise equivalent "Take" treatment. The results show that some Ds take money from Rs in the Take version. Of course, this might be due to Ds, who

in the Give version are otherwise constrained to a corner solution at zero, but that does not explain the lower fraction of Ds who choose positive transfers in the Take version versus the Give version (Bardsley, 2008, List, 2007). This *taking effect* means that the addition of taking options results in less generous givers, indeed, some givers become takers.

From a moral salience approach, it is natural to view giving as moral context, *p*, and taking as non-moral context, *n*. The following assumption fleshes out Assumption 6 for giving and taking.

ASSUMPTION 7: In a dictator game whose Ds and Rs are endowed with *X* and *Y* ($X > Y > 0$), respectively, giving options constitute *p* and taking options *n*. For concreteness, assume the moral measure is $m(C_i) = \max\{c_i \in C_i\} - \min\{c_i \in C_i\}$, $C_i = \{C_+, C_-\}$, where C_+ is the set of non-negative transfers from D to R and C_- the set of negative transfers, i.e., transfers from R to D.

I summarize below many of the rich findings from experiments with taking and propose explanations for them based on moral salience.

SF/Theorem 5.1.1: Consider a standard between-subjects dictator game with endowed Ds and Rs, where $X > Y > 0$. Adding taking options to this game reduces giving on both the intensive and extensive margins, i.e., the mean transfer and the frequency of positive transfers fall (e.g., Cappelen et al., 2013, Cox et al., 2019). The reduction in mean transfers increases, if taking options are enlarged, but less than proportionately (Bardsley, 2008, Korenok, Millner and Razzolini, 2014, List, 2007, Zhang and Ortmann, 2013). The taking effect diminishes, if the D's choice is observable to the experimenter and other subjects (Alevy, Jeffries, and Lu, 2014).

Proof: By Assumptions 6 and 7, adding taking options reduces moral salience, say, from σ^h to σ^l. By Theorem 2.2.1, all dictators for whom $x > 0$ under σ^h, transfer a lower amount under σ^l, with some, who are on the margin, transferring zero or taking. Those, who are constrained at zero under σ^h, take under σ^l. The effect on transfers is less than proportional with taking options from Theorem 2.2.2. The effect of observability follows from Assumption 4 and Theorem 2.2.1.

The results of two studies that vary taking options are presented in panels e and f of Figure 2 (List, 2007, Zhang and Ortmann, 2013, respectively). These illustrate a less than proportionate decline in mean transfers with taking options.

Some dictator experiments vary the endowments of Ds and Rs along with giving and taking options. Specifically, several allow comparisons between a giving game, that is, a standard dictator game where the total endowment, M, is initially all given to the D ($X = M, Y = 0$) and giving is unrestricted, with a taking game, in which M is provisionally allocated to the R ($Y = M, X = 0$) and D taking is unrestricted. Consider now some stylized facts of such games.

SF 5.1.2: In a between-subjects design, where subjects choose under only one condition, R payoffs ($\pi_R = Y + x$) do not differ significantly between the giving and taking games (Chowdury, Jeon, and Saha, 2017, Dreber et al., 2013, Grossman and Eckel, 2015, Korenok, Millner, and Razzolini, 2014, Smith, 2015). In a within-subjects design, where subjects choose under both conditions, R payoffs are lower in the giving game than the taking game, and, given a choice between playing the giving or taking game, most Ds prefer the giving game (86% in Korenok, Millner and Razzolini, 2018).

The utility functions of Ds in the giving and taking games can be written, respectively, as

$U^G = u(M - x) + \sigma \cdot f(\phi(x - \eta)) + \sigma \cdot g(\alpha x)$,

$U^T = u(M - x) + \sigma \cdot f\big(\phi(x - \eta)\big) + \sigma \cdot g(\alpha(x - M))$.

That is, for a given *x*, the D's utility is the same except for the final altruism terms, which reflect the utility from giving from the agent's endowment, in the first case, versus the loss from taking from the patient's endowment, in the second. This leads to the following theorem.

Theorem 5.1.2: In a between-subjects design, it is indeterminate, whether R payoffs will be higher in the giving or the taking game. In a within-subjects design, mean R payoffs are higher in the taking game, and, given the choice between the two games, most Ds prefer the giving game.

Proof: In a within-subjects design, salience is the same for both decisions by Assumption 1. By the concavity of g, $g'(\alpha(x - M)) > g'(\alpha(x))$, at least for altruistic agents, who are assumed to be the average and the majority type ($\mathrm{A}(\overline{\alpha}) - \mathrm{A}(0) > 0.5$), implying a larger x and lower utility in the taking game than the giving game. But in a between-subjects design, salience is lower in the taking game due to the high non-moral context, which we can write $\sigma^T < \sigma^G$, which implies a lower x by Theorem 2.2.1. Thus, the two effects operate in opposite directions in a between-subjects design such that the overall effect on giving is theoretically indeterminate.

Thus, in the absence of a salience effect, transfers should be larger in the taking game. This is, in

fact, what materializes, when salience is the same in the within-subjects design. But when moral salience is lower in the between-subjects taking game, the effect on π_R is ambiguous. Although indeterminacy is a nonspecific prediction, it is not predicted without the effect of salience, and it is consistent with the insignificant differences in π_R in this case compared to the higher π_R in the taking game in the within-subjects design.

Numerous explanations have been offered for the taking effect. Bardsley (2008) conjectures that it is an experimental artefact, viz., an experimenter demand effect, that is, a desire to please the experimenter. In this context, subjects view the offered choice set as signaling what the experimenter wishes the subject to do. But in a recent and rigorous analysis of experimenter demand effects, de Quidt, Haushofer and Roth (2018) find such effects to be modest, and they do not seem plausibly to explain the large magnitude of the taking effect. Cappelen et al. (2013) test whether the choice set signals entitlements, e.g., a taking opportunity might signal the D is morally entitled to do so. But they find that reinforcing entitlements with a real task has no significant effect while the taking effect remains. Korenok, Millner and Razzolini (2012, 2018) point to warm glow and taking aversion (or an endowment effect) and, indeed, altruism in the current model is equivalent to the melding of these two effects. Nevertheless, that alone does not explain the between- versus within-subjects differences. Alevy et al. (2014) argue their results on observability and gender are consistent with social- and self-signaling. As previously discussed, I consider signaling arguments credible, but the present framework is offered as a simple account, which also explains the observability effect in terms of moral salience, specifically, moral proximity.

List (2007) proposes a "moral cost function," which Cox et al. (2019) formalize. They propose and test experimentally a theory with moral reference points that depend on choice sets. Despite differences in theoretical formulation and some differences in predictions, I view the current project as having points in common with Cox et al., which along with Kimbrough and Vostroknutov (2016) and Krupka and Weber (2013), underscore the importance of the changes in agent sensitivity to the violation of moral norms based on differences in choice sets. Whereas these approaches assume certain changes in norms and sensitivity to norms, the present theory derives these and other patterns from a general theory of stable moral norms and context-dependent moral salience.

5.2. Sinners and Saints

This section introduces an experiment that provides an out-of-sample test of the theory presented in this paper while also shedding light on prior findings, especially in relation to the taking effect. In what I will call the "sinners and saints" game, there is a first stage in which dictators may give to or take an amount x from the endowments of the dictators and recipients, where $X > Y > 0$. The endowments are always fixed at the same level, but the range of permissible transfers varies across "*cases*." Cases are varied between subjects, whereby, for a given case, the minimum possible transfer, i.e., the most the D can take, is denoted $x^L \leq 0$, and the maximum possible transfer is denoted $x^H > 0$. Then, there is an unannounced second stage in which a spectator is paid a fixed amount, $\bar{z}$, to allocate an additional larger sum, $Z > X + Y$, between the D and the R in the first stage, which is contingent on each possible D choice for x, i.e., the strategy method is used. To employ a metaphysical conceit, an agent decides how to treat a patient during his mortal life, ignorant of the afterlife in which an impartial judge metes out sanctions involving even higher stakes on sinners and saints. This experiment, while novel, merges design features that have been well validated elsewhere and that, therefore, relate to a broad set of results. The focus of most of the analysis is on second stage spectator decisions in the game described thus far, although the stakeholder decisions in this treatment also serve to replicate prior evidence on the taking effect.[16] The results of this experiment, together with additional decisions reported later that are designed to rule out alternative explanations, support the present theoretical account of previous findings on the taking effect and other anomalies.

The utility function of the spectator in the sinners and saints game is

$$U = u(\bar{z}) + \sigma f\left(\phi\left(z - \eta_z - \theta \cdot r(\hat{\gamma} - \tilde{\gamma}) \cdot {}^{Z}/_{2}\right)\right),$$

where, as in section 4.1, moral preferences are formulated with respect to the D from the first stage, whose choice provides a signal of the D's character. Thus, z denotes the amount of Z that the spectator allocates to D with the remainder of $Z - z$ going to R, η_z is the D's entitlement in the second stage, $\hat{\gamma}$ is D's estimated notional generosity to R, and $\tilde{\gamma}$ is the spectator's threshold

[16] Spectators in a two stage dictator experiment were introduced in Croson and Konow (2009), but that study differed in other ways from the present one: in some treatments, endowments were random and not chosen, in other treatments, only Ds were endowed and could choose transfers but Rs were not endowed, the range of permissible transfers was not varied, and there were never any taking options. As with the current design, Krupka and Weber (2013) use spectators and stakeholders to analyze taking, but their design differs in that third parties are not used to sanction but rather to estimate appropriateness ratings of other subjects on a point scale, stakeholder endowments are varied, and taking options are not varied.

for rewarding or punishing D. The usual assumption that fairness reduces to equal splits in a simple D game like this implies that $\eta_z = {}^{Z}/_{2} - {}^{X}/_{2} - {}^{Y}/_{2} + x$, i.e., η_z calls for equal splits of the total endowments and corrects for any shortfall or excess vis-à-vis equality in D's first stage transfer.

The following theorem states several predictions for this experiment.

Theorem 5.2.1: In the sinners and saints game, second stage spectators sanction, and sanctions are concave in dictator first stage transfers. There is a discontinuous increase in reward (or decrease in punishment) at x^H, and a converse discontinuity at x^L. Holding x^H constant, increasing dictator taking options, i.e., lowering x^L, implies a lower threshold for spectator sanctioning, $\tilde{x}$, and also increases reward, or decreases punishment, of dictators by spectators at every level of dictator transfers.

The proof of this theorem can be found in the Appendix, but the reasoning is illustrated in Figures 6 and 7. First, the context of the reference state may be defined to match the salience in the first stage of this dictator game, $\tilde{\sigma}$. In a game with the same set of permissible transfers as in the reference state, expected generosity, $\hat{\gamma}$, equals notional generosity, γ, both of which equal the dictator's transfer, x, for the full range of dictator types from the least generous, $\underline{\gamma}$, which is the greatest lower bound of notional generosity, to the most generous, $\bar{\gamma}$, which is the lowest upper bound of notional generosity. This is illustrated in Figure 6 by the 45-degree line. Thus, we can write notional generosity, $\gamma(x, \tilde{\sigma})$, as a function of x and $\tilde{\sigma}$. As noted in sections 2.3 and 4.2, however, notional generosity may differ from effective generosity, x, because of constraints on choices. Suppose at least some dictators are constrained to giving less than their preferred amount, i.e., $x^H < \bar{\gamma}$, and/or from taking less than their preferred amount, i.e., $x^L > \underline{\gamma}$. Due to this censoring, a spectator's estimate of the notional generosity of a dictator who chooses x^H, $\hat{\gamma}(x^H, \tilde{\sigma})$, is greater than that of the dictator type, who would choose x^H in the reference state, $\gamma^H = \gamma(x^H, \tilde{\sigma})$, since it includes not only those who notionally prefer x^H but also others who prefer a larger transfer but are prevented from transferring it. Similarly, the spectator's estimate of the generosity of a dictator who chooses x^L, $\hat{\gamma}(x^L, \tilde{\sigma})$, is less than that of the dictator type, who notionally prefers x^L, $\gamma^L = \gamma(x^L, \tilde{\sigma})$.

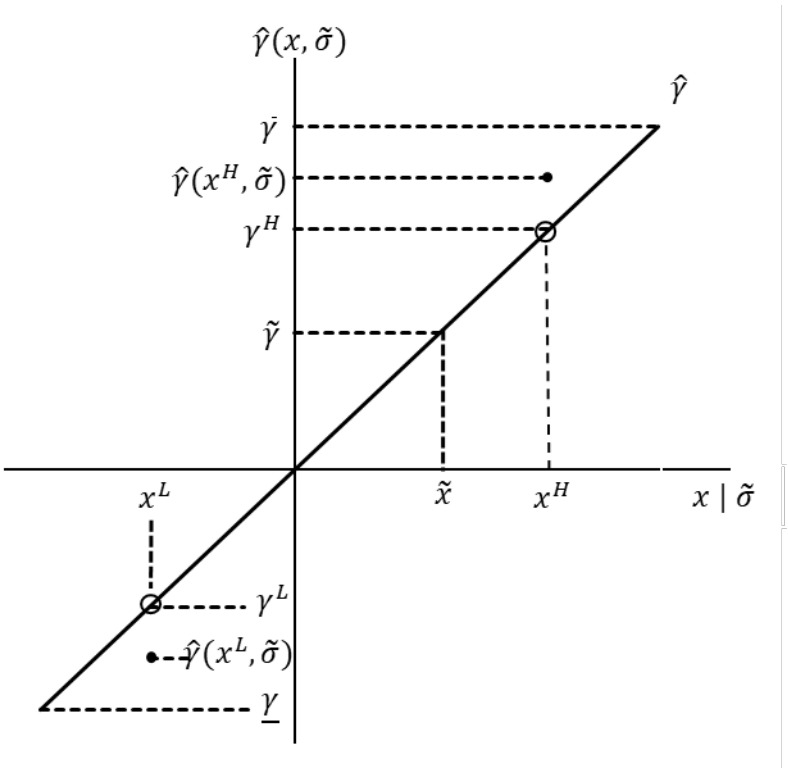

FIGURE 6. – Notional and effective generosity.

Suppose that the spectator's character threshold for dictator generosity is $\tilde{\gamma}$ and that $\gamma^L < \tilde{\gamma} < \gamma^H$. Then, in the interior, $\tilde{\gamma} = \tilde{x}$, and the spectator rewards all $x > \tilde{x}$ and punishes all $x < \tilde{x}$, and, generalizing Theorem 4.1.2, sanctions are asymmetric due to the concavity of r in x. One exception to concavity is occasioned by the censoring of γ at x^H: the optimal sanction, $\tilde{z}$, is increasing in γ, so the discontinuity here implies a discontinuous increase in reward (or reduction in punishment). Thus, the maximum permissible transfer will be treated separately in the regression analysis later.

Other claims of Theorem 5.2.1 are illustrated in Figure 7, which focuses on interior solutions. Estimated (and notional) generosity can be written as a function of x and σ, i.e., $\hat{\gamma}(x,\sigma)$. The main analysis involves variation in the amounts that may be taken, which, as already discussed, affects moral salience. Starting from the reference level of salience, $\tilde{\sigma}$, consider an increase in taking options. Ceteris paribus, salience falls to σ^l, $\sigma^l < \tilde{\sigma}$, and a dictator, who would make a transfer, say, equal to the spectator's threshold of $\tilde{\gamma}$ under $\tilde{\sigma}$, will now give less, $\tilde{x}^l$. That is, the schedule representing the dictator's notional generosity shifts to the left, $\gamma(x,\sigma^l)$, and the spectator's threshold for sanctioning falls (the line going through the origin that corresponds to the reference state is omitted here to avoid clutter). Similarly, a

reduction in taking options increases salience to σ^h, $\sigma^h > \tilde{\sigma}$, and the same dictator will now give more, $\tilde{x}^h$, shifting the schedule to the right, $\gamma(x, \sigma^h)$, and increasing the spectator's threshold for sanctioning. Finally, changes in taking options also affect the spectator's estimate of a dictator's type and, therefore, the spectator's sanctioning of the dictator (which is the same as the dictator's notional type for interior solutions). A dictator, who gives x under low salience, σ^l, reveals higher intrinsic generosity, $\hat{\gamma}^h$, than one, who gives the same amount under high salience, σ^h, and reveals lower generosity, $\hat{\gamma}^l$. Thus, the same behavior may be praiseworthy in the former case but blameworthy in the latter, according to spectators. The final claim of Theorem 5.2.1 concerns the size of the discontinuous increase in z at x^H and the relationship to x^L. An increase in x^L reduces σ, censors fewer dictator types, and increases $\hat{\gamma}(x^H, \sigma)$. This reduces $D(\sigma)$, if the increase in $\hat{\gamma}$ is smaller than the change in censored types.

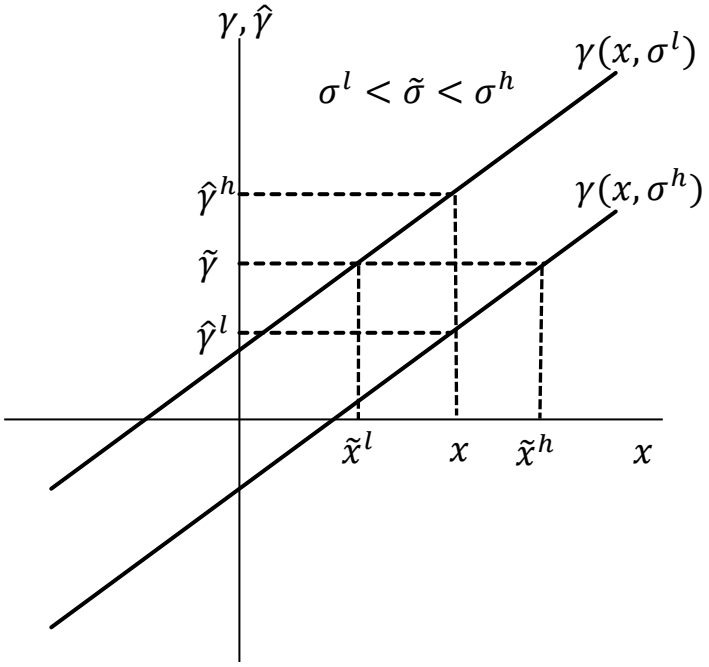

FIGURE 7. – Generosity and moral salience.

The parameters of this experiment are as follows: $X = 15$, $Y = 5$, $Z = 40$, and $x^H = 5$ for all of the three main cases, which differ according to the values of x^L, viz., the Give 5 case with x^L=0, Take 1 case where $x^L = -1$, and Take 5 case where $x^L = -5$. Note that $x^H = 5$ allows X to equalize payments between the X and Y. The values for x^L were chosen to permit testing of the theoretical predictions and to allow comparisons with prior taking games. The main treatment of the sinners and saints game that has been described thus far is called the Double

dictator treatment: subject X chooses how much to transfer to or from subject Y out of their aggregate 20 points, and subject Z is paid a fixed \$5 to allocate 40 points between subjects X and Y for each possible transfer by X to Y, where one point always equals \$0.20. In order to examine whether the entitlement changes with cases, there is also a so-called Benevolent dictator treatment in which X and Y are endowed as in the other treatment, i.e., $X = 15$ and $Y = 5$, and each Z subject is paid a fixed \$2 to choose a transfer of these first stage points between X and Y subjects. Thus, the design is the same as the first stage of the Double dictator treatment with the same cases, except Z instead of X chooses transfers between X and Y, and there is no second stage decision. To avoid any spillover effects of roles and decision contexts on allocations, all decisions were collected between subjects, i.e., separate subjects were used in the roles of X, Y and Z, in the two treatments, and in the different cases. The complete protocol can be found in the online Appendix.

TABLE 1
SINNERS AND SAINTS DESIGN

	Treatment	
Case	Double dictator	Benevolent dictator
Give 5	**66 X**, 66 Y, **66 Z**	30 X, 30 Y, **30 Z**
Take 1	**62 X**, 62 Y, **62 Z**	45 X, 45 Y, **45 Z**
Take 5	**63 X**, 63 Y, **63 Z**	37 X, 37 Y, **37 Z**

The experiment was programmed in oTree (Chen, Schonger and Wickens, 2016) and conducted on Amazon MTurk. Table 1 illustrates the experimental design, showing for each case and treatment the number of participants in each role, whereby those in decision-making roles are denoted in bold font. Similar to many related studies (e.g., Bardsley, 2008, Chowdury, Jeon, and Saha, 2017, Grossman and Eckel, 2015, Korenok, Millner and Razzolini, 2012), a minimum of roughly 30 triples per case were targeted for the Benevolent dictator treatment, and a minimum of twice that number, viz., 60 triples per case, were targeted for the Double dictator treatment, since it is the main treatment of interest. The actual numbers usually exceed these minimums due to differences in the timing of when subjects were cut off from entering. For this study, an MTurk subject pool was preferred for a number of reasons. A substantial literature now exists that MTurk participants behave similarly to university student subjects in qualitative

terms. Moreover, MTurkers are typically closer to the general population in terms of demographic characteristics and average experimental behavior, e.g., Snowberg and Yariv (2021) find the average generosity of MTurk dictators intermediate to that of the more selfish students and that of a more generous representative sample.[17] In addition, the total sample size desired for this study was larger than that accessible at any given time from most student subject pools (the results are based on a total of 1029 participants). Moreover, this study lends itself to the adoption of measures to address typical concerns about an online subject pool (e.g., see Hauser, Paolacci, and Chandler, 2019).[18] Including the $2 show-up fee (called a reward in MTurk), the average earnings were $6.25 for an average of 20-30 minutes of most subjects' time. This is similar to the hourly MTurk wages used by Snowberg and Yariv, which they find is robust to cutting in half and report to be several times the usual MTurk pay of $1-$5 per hour.

TABLE 2
TRANSFERS BY X TO Y IN DOUBLE DICTATOR TREATMENT

	Mean transfer (SD)	Positive transfers (%)
Give 5	3.45 (2.105)	78.8
Take 1	2.16 (2.343)	64.5
Take 5	1.78 (3.777)	61.9

Turning now to the results, Table 2 summarizes the transfers of X subjects in the Double dictator treatment. As predicted, the mean transfers and percentage of positive transfers decrease, as x^L falls. According to two-sided t-tests tests of differences in means, the mean transfer is

[17] Johnson and Ryan (2019) conclude that quality is not harmed by the lack of control and lower stakes on MTurk. Moreover, the equivalency of results from student and MTurk subjects extends to designs involving moral preferences, such as prisoner's dilemmas (e.g., Horton, Rand and Zeckhauser, 2011), public goods games (e.g., Arechar, Gächter, and Molleman, 2018), and dictator games (e.g., Snowberg and Yariv, 2021).

[18] To address concerns about English language fluency, participation was restricted to US residents. Numeracy was established with a test consisting of three fill-in-the-blank questions on addition and subtraction that permitted at most two attempts each before disqualification. To address concerns about attention and comprehension, subjects had to complete correctly within two attempts each of three questions in a quiz about the instructions (two quizzes of three questions each in the case of Z subjects in the more complicated Double dictator treatment). To minimize attrition, each subject faced only one type of decision and the non-strategic design permitted non-simultaneous collection so that subjects did not have to wait for other subjects. The simple design helped keep the study short and address both subject attention and attrition: subjects were permitted up to one hour, but most completed it in less than thirty minutes. Self-selection biases are presumably less problematic with MTurk than with a university subject pool, but that concern was further addressed by describing the study in general terms as an "Academic experiment involving decisions about the distribution of money." As an aside, the data collection took place during the 2020-21 COVID pandemic at a time when laboratory experiments were not feasible, but that fact had no bearing on the choice of an online format, which had been previously planned based on its advantages for this experiment.

lower in the Take 1 (p=0.001) and in the Take 5 (p=0.002) cases than in the Give 5 case, but Take 1 does not differ significantly from Take 5 (p=0.497). Compared to the Give 5 case, two-sided z-tests tests of proportions show a decrease in positive transfers that is marginally significant in the Take 1 case (p=0.073) and significant at conventional levels in the Take 5 case (p=0.036), but Take 1 and Take 5 do not differ significantly (p=0.762). These findings are similar in direction and significance to prior related taking games except that these X subjects are, on average, more generous and less likely to take, when given the opportunity. This is consistent with the expectation that the more representative sample here is more generous than student subjects, who were used in prior studies. A second contributing factor is surely the relatively high ratio of X to Y endowments of 3:1 here versus the lower ones (usually 2:1) used in prior studies, including Bardsley (2008), List (2007), and Zhang and Ortmann (2013).

TABLE 3
MEAN ALLOCATIONS BY Z SUBJECTS TO X SUBJECTS IN DOUBLE DICTATOR TREATMENT

X transfer	-5	-4	-3	-2	-1	0	1	2	3	4	5
Equalizing Z allocation	10	11	12	13	14	15	16	17	18	19	20
Take 5	5.17	5.89	6.03	7.13	8.02	13.08	15.27	16.19	17.92	19.40	20.70
Take 1					6.10	9.53	12.87	15.85	17.73	18.65	21.71
Give 5						7.82	11.44	12.82	13.68	16.03	19.08

Key: Red (blue) allocations are below (above) equalizing ones. Means differ from equalizing transfers according to t-tests at the 5%/10% level of significance; lightly shaded results are not significant at conventional levels.

Having established that X decisions are largely consistent with theoretical predictions for stakeholders and qualitatively replicate prior findings, we turn now to the spectator Z allocations in the Double dictator treatment. Table 3 provides a summary of Z transfers to X subjects. The first row presents the full range of possible X transfers and the second row the corresponding Z transfers to X that equalize total earnings between X and Y. The remaining rows report the mean allocation by Z to X, z, for each level of X transfer, x, which permits a preliminary impression of the results. As predicted, z is monotonically increasing in x within each case in every instance and monotonically decreasing in x^L for each given value of x save in one instance (Take 1, $x = 5$). Comparison of these means with the equalizing Z allocations suggest that Zs usually punish Xs, on average (shaded in red), for transferring less than the 5 points that equalize first stage payoffs: z exceeds the value that equalizes total earnings (shaded in blue), only for an x of 4 or 5 in the Take 5 case and for an x of 5 in the Take 1 case. This is consistent with the predicted shift in $\tilde{x}$ with x^L: the change from Take 5 to Take 1 to Give 5 represents a progressive increase in x^L

and in salience and, therefore, an increase in the threshold for rewarding X transfers.

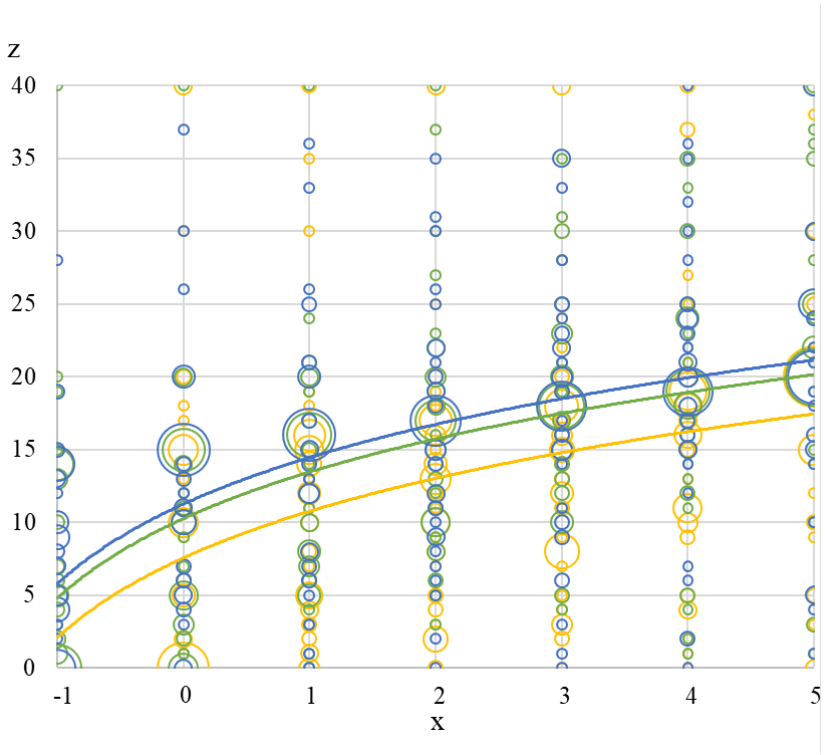

FIGURE 8. – Scatterplot and logarithmic regression trendlines of Z allocations to X (z).
Key: Yellow: Give 5, Green: Take 1, Blue: Take 5

Figure 8 presents a scatterplot of Z decisions, where circle sizes are proportionate to the frequency of each choice and colors correspond to cases: yellow for Give 5, green for Take 1 and blue for Take 5. The main patterns are consistent with theory: z is increasing in x, on average, and some Z subjects equalize, indicated by larger circles along a diagonal, whereas others sanction, reflected mostly by punishment below the diagonal. Turning to multivariate analysis of Z decisions, consider the following regression equation:

$$z_{it} = \alpha + \beta_x \ln(x_t + 2) + \beta_1 Take1 + \beta_5 Take5 + \beta_L x^L + \beta_H x^H + \varepsilon_i$$

where z_{it} is the allocation of Z subject i to subject X based an X transfer of x_t, α is the constant, the βs are the coefficients on the independent variables, and ε_i is the error term. The omitted case is Give 5, so Take1 and Take5 are dummy variables for those cases, respectively. Discontinuities are predicted at x^L and x^H, so dummy variables are included at the lowest possible transfer in each respective case for x^L and at $x^H = 5$, which is common to all cases. The analysis focuses on Tobit regressions with left-censoring using a logarithmic specification of x_t for values between -1 and 5, which is why 2 is added to x_t, making the minimum value of this

independent variable 1. There are several reasons for these choices. First, the non-linear specification is consistent both with theoretical expectations and with the results of regression estimations illustrated by the logarithmic trendlines in Figure 8. Second, the fact that second stage allocations are twice as large as the stakes in the first stage (viz., 40 vs. 20 points) is a feature designed to give wide berth to second stage allocators, and inspection of the scatterplot suggests this was largely successful. Nevertheless, scatterplots also reveal censoring of z at lower values of x and recommend the use of Tobit. Third, the chief interest is in comparison of differences across cases where there are some common values of x for the cases, i.e., from -1 to 5. Moreover, including additional negative x values for the Take 5 case produces results that are qualitatively similar but risks producing estimates that give disproportionate weight to the Take 5 case and its increasingly censored values (reaching 46% of allocations when x equals -5).

TABLE 4
REGRESSION ANALYSIS OF Z ALLOCATIONS TO X

	(1) OLS	(2) Tobit
$\ln(x+2)$	6.516***	6.822***
	(0.509)	(0.352)
Take 1	2.507*	2.662*
	(1.105)	(1.112)
Take 5	3.587**	3.680***
	(1.169)	(1.114)
x^L	−0.580	−1.533*
	(0.749)	(0.636)
x^H	1.769***	1.673***
	(0.417)	(0.464)
Constant	4.021***	3.398***
	(1.188)	(0.941)
R^2	0.255	

N=191. Tobit regressions are left-censored. Standards errors are clustered at the individual subject level and are reported in parentheses. */**/*** denotes significance at the 5/1/0.1-percent level.

Table 4 presents the results of regression analysis, whereby standard errors are clustered at the level of the 191 individual Z subjects. Column 1 shows the results of an OLS regression and column 2 the results of a Tobit regression with left-censoring. The results are the same except for the x^L dummy variable, which is negative in both estimations but turns significant when account is taken of the left-censoring that is compromising the OLS estimation. We focus, therefore, on regression (2), the results of which are all consistent with Theorem 5.2.1, some at

high levels of significance. Spectator sanctions are increasing and concave in X transfers to Y. As taking options expand, spectators progressively increase reward, or decrease punishment, significantly in the Take 1 and Take 5 treatments. The coefficient for the x^L dummy confirms the predicted discontinuous decrease in z and the coefficient for the x^H dummy the predicted discontinuous increase in z.

The theory advanced here has been shown to be highly consistent with the results on stakeholder and spectator decisions in the sinners and saints game. As always in such instances, that fact does not rule out the possibility of alternative explanations, such as those offered by the reciprocity theories mentioned in section 2.3. For example, Rabin (1993) and Dufwenberg and Kirchsteiger (2004) define fairness as the average of the highest and lowest efficient payoffs, which implies fair allocations (η) vary directly with x^L. Falk and Fischbacher (2006) define fairness as equal payoffs, but their theory generates equivalent predictions about the effect of variation in x^L by building the effect of the choice set into their "intention factor." Reciprocity theories are typically formulated for stakeholders, but, for simplicity, I will cast them in the current spectator framework and analyze the effect of x^L through η.

Theorem 5.2.2: In the sinners and saints game, let the fair allocation in the first stage and the threshold for sanctioning be functions of the minimum permissible transfer, i.e., $\eta_x(x^L)$ and $\tilde{x}(x^L)$, respectively, where $\frac{\partial \eta_x}{\partial x^L} > 0$ and $\frac{\partial \tilde{x}}{\partial x^L} > 0$. Then lowering x^L increases reward, or decreases punishment, at every level of dictator transfers.

Proof: See Appendix.

Reciprocity theories provide a partial account for observed spectator sanctioning: they are consistent with the observed shift in the threshold and the level of sanctioning. They do not, however, predict the discontinuities we observe at x^L and x^H. Moreover, reciprocity theories and the theory proposed here diverge in their predictions for behavior in the aforementioned Benevolent dictator treatment, in which a spectator chooses transfers between X and Y. In this treatment, the spectator's utility function is simply

$$U = u(\bar{z}) + \sigma f\big(\phi(x - \eta_x)\big),$$

which yields the following predictions.

Theorem 5.2.3: In the Benevolent dictator treatment, the spectator allocates to subjects their entitlements. That means the spectator's allocation does not vary with x^L according to the

theory of moral salience, conditional altruism and virtue preferences, but it does vary directly with x^L, according to reciprocity theories.

Proof: By the first order condition, $\sigma f'\big(\phi(x-\eta_x)\big)=0$, which implies $x=\eta_x$. In this game, η_x is fixed in the theory advanced in this paper, whereas $\frac{\partial \eta_x}{\partial x^L}>0$ in reciprocity theories.

TABLE 5
TRANSFERS BY Z TO Y IN BENEVOLENT DICTATOR TREATMENT

			Difference in Means p-values (t-statistics)			
	Mean transfer (SD)	*N*	$H_0: \eta_x = 5$	Give 5	Take 1	Take 5
Give 5	4.47 (1.335)	30	0.034 (2.175)			
Take 1	4.04 (1.673)	45	0.000 (3.849)	0.242 (-1.179)		
Take 5	4.41 (1.077)	37	0.001 (-3.332)	0.839 (-0.204)	0.249 (1.161)	
Give 10	4.68 (1.738)	40	0.245 (1.165)	0.584 (0.551)	0.088 (1.729)	0.420 (0.812)

Table 5 presents the results of this treatment for the three main cases as well as for an additional case, Give 10, which I will discuss momentarily. For the three main cases, the mean transfers range from \$4.04 to \$4.47. The three pairwise tests reported in the table reveal that none of these differences is significant even at the 20% level, whereby all p-values in this table are two-sided. This supports the theory proposed here, which assumes a fixed entitlement, over reciprocity theories, in which it changes. A different question concerns spectators' estimate of what entitlement is fair. The theoretical analysis has often proceeded from the assumption that the entitlement in simple games like these equalizes earnings, i.e., for this decision, that $\eta_x = 5$. But the tests reported in the Table 5 show that the means for the three main cases all differ significantly from 5. The assumption of equalizing entitlements is one of convenience rather than necessity, since results rarely depend qualitatively on whether η_x equals 5 or 5 minus some epsilon. Nevertheless, while comparisons of mean spectator transfers provide valid conclusions about whether the entitlement varies with x^L, they might not provide good estimates of the entitlement itself. As with all experiments, subject decisions here are noisy, but variance is censored on the right at the value of 5 in the three main cases, which creates a downward bias in the estimate of η_x. For that reason, the benevolent dictator treatment includes an additional case,

Give 10, in which spectators may allocate any amount from 0 to 10 to subjects X. For this case, the mean transfer is $4.68, which does not differ at conventional levels from the other cases, and, to the point, does not differ significantly from 5.

This section introduced an experiment designed as an out-of-sample test of the theory proposed here. The results of the experiment are uniformly consistent with the predictions of the theory and, in several respects, inconsistent with reciprocity theories. Below we continue to apply the theory to additional decision contexts.

5.3. Joy of Destruction

People sometimes destroy the wealth of others at a personal cost, often risking punishment, and with no material benefit to themselves, e.g., some people vandalize property or write computer viruses. There are examples of people, who cooperate over generations but suddenly begin engaging in destructive behavior toward one another. For instance, Serbs, Croats and Bosniaks lived peaceably, often intermarrying, prior to the breakup of Yugoslavia, but subsequently turned on one another, and over 100,000 lives were lost and vast amounts of property destroyed in the Bosnian War. Depending on the particular case, such behavior might be attributed to ethnic hostility, preemptive retaliation, revenge, etc. Economics experiments, however, have documented that, even when such motives can be ruled out by design, some people are willing to incur a cost to destroy the wealth of others and that such behavior can be easily triggered.

Various "money-burning" games have found that up to nearly one-half of subjects acting individually in simultaneous games with unequal endowments destroy earnings of other members of their group, e.g., Zizzo and Oswald (2001), Abbink and Sadrieh (2009), Abbink and Hermann (2011). Similar behavior is observed, when players interact over multiple periods in so-called "vendetta" games, e.g., Abbink and Herrmann (2009), Bolle, Tan and Zizzo (2014). In these studies, however, one cannot rule out motives other than a pure desire to destroy. When endowments are unequal, subjects can be motivated by inequality aversion to destroy. Moreover, as we will see, relatively few subjects destroy in non-strategic decisions (13-15%), but a much higher percentage expect others to destroy (38% in Abbink and Herrmann, 2011), which is consistent with preemptive retaliation in these experiments. Thus, we will focus, as usual, on simple, non-strategic decisions in the cases that follow, such as non-strategic versions of the

"joy-of-destruction" (or JD) game, which resembles a dictator game, in that it is unilateral, but with options to destroy others' endowments. In the standard version, endowments are equal, and agents can destroy at zero cost, and zero benefit, to themselves.[19] The utility function of the agent in a non-strategic JD game can be written

$$U = u(X) + \sigma \cdot f(\phi(Y + x - \eta)) + \sigma \cdot g(\alpha x).$$

In the standard JD game, $x \leq 0$, the context is simple and endowments are equal, so we assume $X = Y = \eta = M/2$, where $M = X + Y$. We will also consider cases where endowments are unequal and the agent may destroy or create money for the patient, i.e., x can be positive or negative.

SF 5.3.1: In the standard non-strategic JD game with symmetric endowments, a minority of agents engages in destruction (13% in Iriberri and Rey-Biel, 2013, 15% in Kessler, Ruiz-Martos and Skuse, 2012, and only 13% even in the strategic game of Abbink and Herrmann, 2009). In a JD game where $X > Y$, allowing agents not only to destroy but also to create reduces the fraction who destroy and increases the average transfer (Zhang and Ortmann, 2013).

Theorem 5.3.1: In the standard non-strategic JD game with $X = Y$, only a minority, consisting solely of spiteful agents, destroys. Adding creation to a JD game where $X > Y$ reduces the fraction that destroys and increases the average *x*, if the addition of creation affects altruism salience more than fairness salience.

Proof: See Appendix.

If only spiteful subjects destroy when $X = Y$, SF 5.3.1 implies 13-15% of subjects in those studies are spiteful. Moreover, adding positively valenced opportunities to create wealth to a JD game reduces the frequency and mean level of destruction assuming the type of salience affected is mostly, or solely, altruism salience.

SF/Theorem 5.3.2: When endowments are unfair, destruction is directed mostly toward those with unfairly high endowments. That is, destruction is largely toward richer subjects in games with unearned endowments (Zizzo, 2003, Zhang and Ortmann, 2013), but, with earned endowments, destruction is directed toward richer subjects for the most part only

[19] Thus, the agent's endowment is fixed in most JD games, as we assume in the current analysis. Although it shares this feature with spectator decisions, this is, nonetheless, treated as a stakeholder decision and, therefore, the altruism term is included in the agent's utility function. This is because the JD context casts the agent in a personal, agent-patient relationship, similar to a dictator game, and not as a spectator choosing impartially for others.

if inequalities are unfair (Fehr, 2018).

Proof: See Appendix.

These findings underscore the importance of inequity aversion, as opposed to spite alone, in explaining much of the destruction in JD, money burning and vendetta games such that even some altruistic subjects engage in destruction.

6. Norm Avoidance

This section deals with three anomalies that have been studied extensively. They involve contexts that provide opportunities for agents to reduce the salience of moral norms. Just as some of the helping and harming anomalies in section 5 related more to variation in altruism, it is natural for some findings discussed in this section to rest on the comparatively greater importance of differences in fairness preferences, given the focus here on norms. One may think of fairness salience as being more important than altruism salience in these phenomena, although that is not a necessary assumption for any claims here.

I discuss three types of norm avoidance, which involve, respectively, exiting a situation with a choice that is high in moral salience, avoiding information that helps identify the moral but personally costly choice, and delegating the choice to others. Assumption 8 expresses more explicitly statements in section 2.2 about moral salience and applies them to the avoidance of moral norms.

ASSUMPTION 8: Compared to moral salience in the standard dictator game (σ^h), moral salience is lower with the availability of an option to avoid taking action on or acquiring information about the consequences of one's action (σ^m), even if the agent does not exercise that option. Moral salience is lower still for those who actually exercise the option and choose to avoid the action or information about the consequences of the action (σ^l).

That is, the option to avoid a moral norm, or information about its consequences, is non-moral context, which lowers moral salience. Actually exercising that option lowers moral proximity and, therefore, moral salience further. These points are fleshed out below for each of the three cases in this section.

6.1. Moral Egress

Some people will give money to a beggar but prefer to cross the street, if possible, to

avoid the beggar. Field experiments have established that avoidance of this kind is widespread. Forewarning people of door-to-door charitable solicitations results in a large and significant drop in the fraction of homeowners, who open their door at the pre-announced hour, compared to those who are not forewarned and who give more generously and at a higher rate (DellaVigna, List, and Malmendier, 2012). Placing Salvation Army bell-ringers at both of two entrances to a supermarket, rather than just one that can be avoided, increases both the rate and level of donations (Andreoni, Rao, and Trachtman, 2017). Laboratory experiments have found these results to be robust to controls for possible extrinsic motives, such as social pressure or social image concerns. I will call this anomaly *moral egress*: people comply with moral norms, when the norms are salient and exit is prohibitively costly or impossible, but many prefer to exit a situation with high moral salience, when possible.

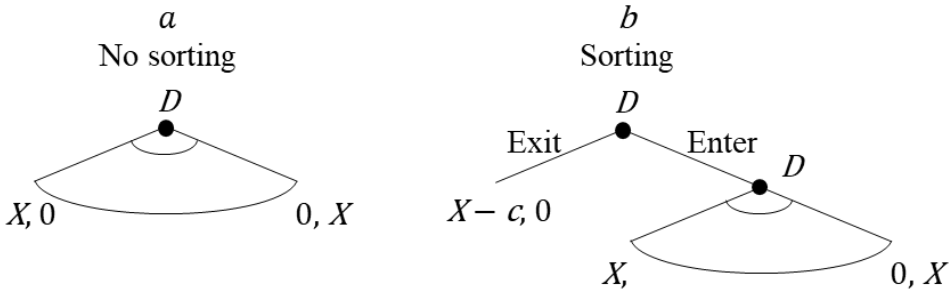

FIGURE 9. – Exit game of Lazear, Malmendier and Weber (2012).

Dana, Cain and Dawes (2006) introduced an experiment in which dictators first play a standard dictator game with $10. Then the Ds are told for the first time that they may either implement the division or exit the game with $9, in which case the Rs receive nothing and never find out about the D decision. In the standard game, mean D transfers are at the usual level of about 25% of the endowment, but up to 43% of Ds choose instead to exit and take the $9. Exit is inconsistent with standard social preference models: selfish Ds should stay in the game and take $10 instead of $9, whereas fair-minded Ds should also stay but share fairly. The discussion here centers on a version of this experiment by Lazear, Malmendier and Weber (2012) that extends the Dana et al. design and lends itself to further analysis. This version, which I will call the "exit game," is illustrated in Figure 9. Panel a shows their "no sorting" treatment, which is a standard D game, and panel b illustrates a "sorting" treatment. In the sorting treatment, Ds first choose whether to enter or to exit the D game. They know that, if they enter, they will then proceed to make a decision about how much to transfer of their endowment, X. If they choose to exit, they

receive X less a cost c, where $0 \leq c < \eta = \frac{1}{2}X$. That is, in some variations, the exit cost is zero, but, when positive, it is smaller than the common entitlement, which I take to be equal splits in this simple D game.

Consider the following stylized facts from some experiments with exit options.

SF 6.1.1: In dictator games with exit, some dictators enter and transfer zero, some enter and transfer a positive amount, and some exit. Those who exit are, on average, more generous, than those who enter (Broberg, Ellingsen, and Johannesson, 2007, Dana, Cain and Dawes, 2006, Lazear, Malmendier, and Weber, 2012).

For the analysis, Assumption 8 means that, relative to moral salience in the standard dictator game with no exit (σ^h), the availability of an exit option lowers moral salience for those, who elect to enter (σ^m), and moral salience is even lower for those who exit and avoid the dictator decision altogether (σ^l). To simplify the analysis, I assume, without loss of generality, that $\sigma^h = 1$, $0 < \sigma^m \equiv \sigma < 1$, and $\sigma^l = 0$. Denote the utility of a D who enters the game UN, one who exits UX, and one in a standard dictator game without any exit option UD. Then, their utility functions are

$$UD = u(X - x) + f(\phi(x - \eta)) + g(\alpha x).$$

$$UN = u(X - x) + \sigma f(\phi(x - \eta)) + \sigma g(\alpha x),$$

$$UX = u(X - c),$$

The following theorem addresses entry, exit and ranges of transfers in the exit game.

Theorem 6.1.1: In the exit game, some dictators enter and some exit. Of those who enter, some transfer nothing and some make a positive transfer. The utility of entry, and its attractiveness relative to exit, is increasing in α (for $x > 0$) and decreasing in ϕ (for the mean and majority of dictators for whom $x \leq \eta$). If variation in generosity in the exit game is due mostly to variation in fairness preferences, then more generous dictators prefer, on average, to exit, i.e., those with fairness coefficients above a threshold ϕ^X.

Proof: See Appendix.

A range of optimal transfers among those who enter follows from previously assumed differences in generosity in dictator games. The utility of entering varies with the strength of altruism and fairness preferences, and the fact noted in SF 6.1.1 that more generous Ds are more likely to exit is consistent with generosity in this game being driven chiefly by differences in fairness preferences, since fairer Ds prefer to exit.

The following stylized fact concerns mean transfers.

SF 6.1.2: In a dictator game with exit, mean transfers are lower than in the standard dictator game without exit (Broberg, Ellingsen, and Johannesson, 2007, Dana et al., 2006, Lazear et al., 2012).

The theory predicts this SF for the different designs with exit, but it is worked out formally for the exit game in the following theorem.

Theorem 6.1.2: In the exit game, mean transfers are lower among those who enter than in the standard dictator game without exit.

Proof: Moral salience is lower in the exit game than in the game without exit, so the optimal transfer x is lower by Theorem 2.2.1 combined with SF 6.1.1 that more generous Ds exit at higher rates.

Other findings from games with exit can be explained by changes in moral salience. Lazear et al. (2012) and Andreoni et al. (2017) find that exit increases, if agents must confront patients face-to-face, and Dana et al. (2006) find that both exit and transfers fall, if Rs are never told there was a dictator game regardless of whether the D exits. As described in section 3.2, such procedural differences are expected to affect moral salience in the form of moral proximity: moral salience increases with personal knowledge of the agent and even of the existence of the agent in a capacity that can affect the patient.

Experiments with exit have produced other findings that are consistent with this model. One example comes from Lazear et al. (2012), who report the following result from an exit game.

Theorem 6.1.3: In an exit game with a constant value of exit ($\bar{X} - c$), the frequency of exit decreases as the stakes of entering the game (X) increase.

Proof: See Appendix.

There are explanations other than moral salience for some of the experimental results on exit. For instance, if the D chooses to exit, Rs do not find out about the dictator game in these studies, which raises the question of whether Ds are responding to other forces such as social image concerns or guilt aversion, i.e., disutility from giving less than what the R expects. On the latter effect, the evidence is mixed, e.g., Charness and Dufwenberg (2006) find support for guilt aversion in a strategic game with communication, whereas Ellingsen et al. (2010) find the effects are close to zero in three games, including the dictator game. Regarding the former effect, I am

unaware of any tests of social image in experiments with exit, so one cannot rule it out. The relative strengths of the moral egress argument are its simplicity and its position in the broader theoretical framework of moral salience. It rests on the intuition that people sometimes distance themselves from situations in which moral norms are salient, because norm compliance reduces utility. In fact, to the extent utility can be thought of as subjective well-being, there is evidence suggesting that Ds, who are paired with Rs but given no opportunity to share their endowment with them, are happier than Ds in a standard dictator game (Konow, 2010).

6.2. Willful Ignorance

An important factor contributing to the 2007-08 financial crisis was the ability of lenders to avoid documenting applicants' incomes and, thereby, avoid knowing about the borrowers' often inflated claims. In the accounting sandals of the early 2000s, CEOs of troubled firms like Enron and Worldcom later professed ignorance of the dubious accounting practices at their companies. As these examples illustrate, the economic consequences of such information avoidance can be staggering, but this phenomenon is widespread and spans other important domains. Political polarization, for example, can be traced to the shunning of broadcast and online news sources that uncomfortably challenge one's preconceptions (Dahlgren, Shehata, and Strömbäck, 2019, Peterson, Goel, and Iyengar, 2019). People often avoid information that could help them better identify the moral, right or socially beneficial course of action, which I will call *willful ignorance*. Experiments indicate that this behavior has an intrinsic component that cannot be explained, at least not solely, by extrinsic motives such as unadulterated greed, fear of legal culpability, or information costs.

Consider the binary dictator game introduced by Dana, Weber and Kuang (2007), or DWK, which I will call the "information game." Most subsequent studies of willful ignorance in experimental economics employ this design or ones very close to it, although quite different designs have also come to similar conclusions, e.g., Serra-Garcia and Szech (2019) and Spiekermann and Weiss (2016). There are three possible payoffs, H, F and L, that can be paired between D,R in four ways, $\{\pi_D, \pi_R\}$, where $0 \leq L < F < H$ and $L + H < 2F$ ($H = 6$, $F = 5$, $L = 1$ in DWK). The sequence of decisions and payoffs are illustrated in Figure 10. There are two states of the world, $\omega \in \{1,2\}$, that occur with equal probability, $q = 0.5$. The D first chooses whether to reveal the realization of the gamble (R) or not (NR) and then chooses either

option A or option B. If D chooses reveal, D then finds out whether the fairer option is A or B before choosing: in state 1, this is 1B, and, in state 2, this is 2A. As usual in a simple dictator game, fairness reduces to equality, specifically in the information game, I assume the entitlement is the patient's payoff in the fairest possible state, which is F in this game. If D chooses not to reveal, option A or B is chosen without knowledge of the realization of the gamble, whereby the fairer option is B in expectations. Since revealing is costless, though, D can always guarantee the fairer outcome, although that would mean sacrificing a payoff of H should state 1 obtain. There is also a baseline treatment for comparison in which D chooses only between 1A and 1B, the payoffs of which are known to D.

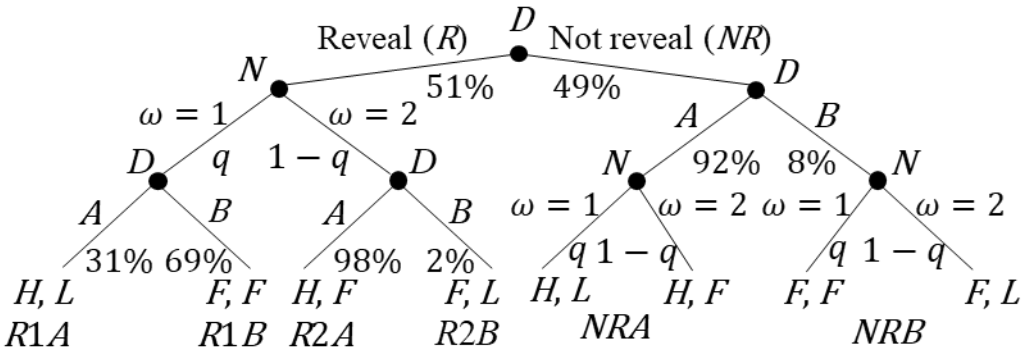

FIGURE 10. – Information game of Dana, Weber and Kuang (2007).

Consider the following stylized facts for this game. The percentages cited are the unweighted averages from DWK plus five other studies that employ this design: Bartling, Engl and Weber (2014), Feiler (2014), Grossman (2014), Grossman and van der Weele (2017), and Larson and Capra (2009). The various studies are quite consistent regarding the following patterns. These percentages as well as the acronyms for the various choices are summarized in Figure 10.

SF 6.2.1: In the information game, Ds are roughly equally split between those who reveal (51%) or those who do not (49%). Of those who reveal, a majority (69%) chooses the fair option B, if state 1 obtains, and nearly all (98%) choose the fairer option A, if state 2 obtains. Of those who do not reveal, almost all (92%) choose the less fair option A. In the Baseline, a majority (71%) chooses the fair option B.

Let the D's choices of A or B in state ω be written $x_\omega = (x_1, x_2)$, where $x_i \in \{A, B\}$. Then the D's utility knowing the realization of the state can be written

$$U(x_\omega, \omega) = u(\pi_D | x_\omega, \omega) + \sigma f(\phi(\pi_R - \eta | x_\omega, \omega)) + \sigma g(\alpha \pi_R | x_\omega, \omega).$$

The D's expected utility before knowing the state can be written

$$EU(x_\omega, \omega) = 0.5 \cdot [u(\pi_D|x_1, 1) + u(\pi_D|x_2, 2)] + 0.5 \cdot \sigma[f\big(\phi(\pi_R - \eta|x_1, 1)\big) + f\big(\phi(\pi_R - \eta|x_2, 2)\big)] + 0.5 \cdot \sigma[g(\alpha\pi_R|x_1, 1) + g(\alpha\pi_R|x_2, 2)].$$

Applied to willful ignorance, Assumption 8 means that the option in the information game to remain ignorant of the consequences of one's choices lowers moral salience, even for those who reveal (σ^m), compared to the standard game (σ^h), and moral salience is even lower for those who do not reveal (σ^l).

The following theorem reconciles the patterns of SF 6.2.1 with moral salience. The proof can be found in the Appendix.

Theorem 6.2.1: In the information game, suppose the effect of options on altruism is second order, i.e., the differences in utility terms between options is smaller for altruism than for material utility and, if applicable, inequity aversion. Then reveal and option B in state 2 (R2B) and not reveal and option B (NRB) are strictly dominated (this need not hold, if altruism is not second order and, specifically, if D is extremely spiteful). The fairest Ds choose to reveal and then option B in state 1 (R1B) and option A in state 2 (R2A). Less fair Ds choose not to reveal and then option A (NRA), except for the least fair Ds, who choose to reveal and option A in states 1 and 2 (R1A and R2A). The percentage of Ds choosing fair in the baseline is greater than the estimated fraction of R1B2A and less than the sum of the fractions choosing R1B2A plus NRA in the information game.

Proof: See Appendix.

According to the theorem, two choices are strictly dominated, viz., not reveal and option A, and reveal and B in state 2, which is consistent with the nearly zero incidence of these choices in the experiments.[20] The fairest Ds want to reveal and choose the fairest options, viz., B in state 1 and A in state 2. The least fair Ds choose to reveal and guarantee themselves the highest payoffs by choosing option A in both states. In between are the Ds, who seek to lower their inequity aversion through reduced moral salience by not revealing the consequences of their choices, i.e., choosing not reveal and A. Finally, the 71% choosing 1B over 1A in the standard game lies in the approximately 35% to 80% range predicted by the theorem for the information game, inferring types under reveal from the frequency of choices of 1A and 1B.

[20] The theorem states this need not hold, if Ds are extremely spiteful: this is noted, not because these are predicted to occur, but because of later claims about how agents would respond to them, if they did occur.

Bartling, Engl and Weber (2014, or BEW henceforth) adapted the DWK design, adding a baseline to state 2 and adjusting the payoffs while maintaining the fundamental relative relationships between L, F and H. They also added third party punishment of the D using the strategy method and based on D choices to reveal or not and of option A or B as well as of the realized payoffs. Below I analyze third party punishment in the information game assuming non-strategic behavior, e.g., punishment by unannounced spectators. I note two differences from the design of BEW. First, the fact of the possibility of punishment was common knowledge in BEW, so that could prompt strategic self-interest by first stage Ds and, thereby, distort the punishers' estimates of the Ds' character. Second, the third parties in BEW were not true spectators, since they paid a price of 0.2 per unit punishment. These two facts can be expected to affect the degree of punishment, but they would still not impact the predictions qualitatively, as long as strategic self-interest does not undo the ranking of character types. The positive price of punishment might only affect the quantity of punishment, and, even so, it is relatively small price.

The following theorem states predictions about spectator punishment in the information game.

Theorem 6.2.2: Suppose altruism is second order, and the mean threshold for generosity is between the estimated level for those choosing fair (1B) in the standard game and for those choosing not reveal and A (NRA) in the information game. Then R1B and 1B will not be punished. There is punishment for all other decisions, specifically, in ascending order of punishment, for NRA in state 2 (NRA2), NRA in state 1 (NRA1), R1A and 1A. Punishment for R2A is greater than 2A, which lie between that for R1B and R1A. Other choices are predicted to be dominated and could only occur, if altruism were not second order and Ds were extremely spiteful. If spectators, nevertheless, sanction for the possibility of such extremely spiteful Ds, and assuming the estimated character of such Ds fall below the character threshold, these choices would be punished in ascending order as NRB in state 1 (NRB1), NRB in state 2 (NRB2), R2B and 2B.

Proof: See Appendix.

To summarize, this theorem makes the following predictions about the ideal punishment $\tilde{z}$ (expressed in non-positive terms) in the information game based on conclusions that can be drawn from the proofs for Theorem 6.2.1 and on the specification of virtue preferences under uncertainty in section 4.2, including the adjustment of the scale.

$$0 = z^{R1B} = z^{1B} > z^{NRA2} > z^{NRA1} > z^{R1A} > z^{1A}$$

$$0 = z^{R1B} > z^{2A} > z^{R2A} > z^{R1A}$$

$$0 > z^{NRB1} > z^{NRB2} > z^{R2B} > z^{2B}$$

Table 6 presents the results from BEW on mean punishment points by choice and, where outcomes are random, realization. The levels of punishment follow exactly the patterns predicted by the theory. In the top rubric, punishment in R1B and 1B is effectively zero, and punishment increases monotonically from NRA2 to 1A. In the bottom rubric, punishment in R2A is greater than in 2A and lies between R1B and R1A with its greater proximity to R1B consistent with the larger predicted share of fair types in R2A. Finally, certain choices are predicted to be dominated in this game, based in part on the assumption that altruism is second order. In fact, these actions are chosen either never by Ds or at frequencies so low that they are consistent with the kind of noisy choice typically observed in experiments, viz., NRB (0% in BEW, 8% in the six studies overall), R2B (0% BEW, 2% overall), and 2B (4% BEW). Thus, these results bolster the assumption that altruism is second order in this design. Nevertheless, it is not contradictory for spectators, who are punishing for all possibilities using the strategy method, to take account of the possibility of Ds, whose spite is extreme and punish choices that, in practice, turn out to be dominated. In fact, the punishment of extremely spiteful types, or of the possibility of such types, also follows the predicted pattern, increasing from NRB1 to 2B.

TABLE 6
PUNISHMENT POINTS IN BARTLING, ENGL AND WEBER (2014)

R1B	1B	NRA2	NRA1	R1A	1A
-0.58	-0.56	-8.00	-11.42	-16.25	-19.72
2A	R2A	NRB1	NRB2	R2B	2B
-1.76	-2.67	-4.42	-6.00	-9.50	-12.41

Numerous explanations have been offered for information avoidance, e.g., see the excellent review of Golman, Hagmann and Loewenstein (2017). On the more specific topic of willful ignorance, which as used here involves a connection to moral preferences, Gino, Norton and Weber (2016) explain willful ignorance based on motivated reasoning, i.e., ignorance allows selfish dictators to believe they are being moral. This seems consistent with other evidence of self-serving fairness biases, e.g., Babcock, Loewenstein, Issacharoff, and Camerer (1995),

Konow (2000), although I am unaware of any evidence on the incentivized elicitation of moral beliefs in the specific case of willful ignorance. Along other lines, Grossman and van der Weele (2017) propose a theory of self-image that is consistent with D behavior in the information game. They argue persuasively in favor of an explanation based on self-signaling, although it is unclear how that theory might explain patterns of third-party punishment. As I see it, though, the main arguments for moral salience are its parsimony and broad range of applications, while being potentially complementary to, rather than conflicting with, alternative explanations such as motivated reasoning and self-image.

6.3. Delegation

Numerous management consulting firms exist largely to recommend or carry out the firing of the employees of their client companies, even though those companies could implement the firings themselves and, thereby, save themselves the consulting fees. Companies in developed nations outsource much of their manufacturing to companies in less developed countries where labor standards are lower, even though there are, in some cases, be cost advantages from vertically integrating foreign production. When decision-makers delegate such choices, it raises the question of whether they seek to deflect blame from themselves for undesirable consequences, say, from their personal involvement in firings or from dangerous work conditions, such as those that led to the collapse of the Rana Plaza textiles factory building in Bangladesh in 2013 that killed more than 1100 workers. In fact, economics experiments corroborate the desire of agents to delegate immoral choices to others after ruling out other reasons, including liability concerns, the value of outside expertise and advantages of specialization.

Experimenters have studied delegation chiefly using dictator games in which a dictator may delegate to an intermediary (I) the decision about the payoffs of the D, R(s), and I. There have been wide variations, though, in features of the designs, such as continuous or binary choices, single shot or multiple rounds, communication between subjects or not, differing numbers of subjects in groups, fixed matching or rematching of groups, selection of Is, different opportunities for punishment and of different members of groups, etc. Nevertheless, certain patterns are robust across these designs and are summarized in the following stylized facts.

SF 6.3.1: When dictators have an option to delegate, average allocations to recipients are lower than in a standard dictator game. Some dictators delegate the allocation decision to intermediaries, who usually choose unfair allocations, and fewer dictators make fair allocations directly themselves (Hamman, Loewenstein and Weber, 2010, Coffman, 2011, Bartling and Fischbacher, 2012, Oexl and Grossman, 2013).

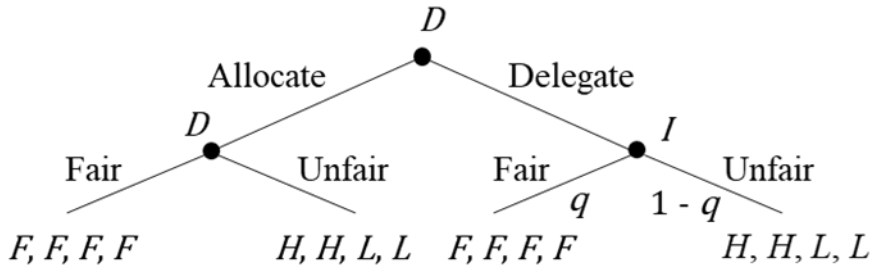

FIGURE 11. – Delegation game of Bartling and Fischbacher (2012).

For the analysis, I focus on a simple, non-strategic design, which I will call the "delegation game," that was introduced by Bartling and Fischbacher (2012, henceforth BF) and that has been used and adapted by others, e.g., Oexl and Grossman (2013). Each group of four consists of a D, an I, and two Rs, whose payoffs, $\{\pi_D, \pi_I, \pi_{R1}, \pi_{R2}\}$, can be fair, $\{F, F, F, F\}$, or unfair, $\{H, H, L, L\}$, where $0 \leq L < F < H$ and $L + H = 2F$ ($H = 9$, $F = 5$, $L = 1$ in BF). The sequence of decisions and payoffs are illustrated in Figure 11. The D chooses either to allocate directly or to delegate the decision to an I, whereby the D is assumed to estimate the probability that the I will allocate fairly to be q, where $0 < q < 1$. There is also a baseline treatment with no delegation option that corresponds to the Allocate branch of the game tree. Clearly, standard theory predicts that a risk neutral or risk averse D will never delegate, since allocating directly guarantees the fair or unfair outcome preferred by D. BF include treatments with punishment that produce high rates of fair choices, likely motivated by strategic self-interest. But in their non-strategic treatments of the delegation game without punishment, 66% of Ds choose unfair, 17% choose fair, and 17% choose to delegate, whereby 82% of Is then choose the unfair allocation. In the baseline dictator game, 65% of Ds choose unfair, almost identical to that in the delegation game, but that means that the percentage of Ds choosing fair directly drops from 35% in the baseline to only 17% in the delegation game. Thus, the delegation option appears to lead about one-half of otherwise fair Ds to delegate.

Applied to the delegation game, Assumption 8 means that the very existence of an option

to delegate the decision in the delegation game to an intermediary lowers moral salience (σ^m) relative to the standard dictator game (σ^h), and moral salience is lower still for those who actually choose to delegate (σ^l). The moral salience of the intermediary, who allocates, is open to some interpretation: since I is aware of the delegation option, salience should be no greater than σ^m, but the fact that D actually exercised that option might lower salience further to σ^l, hence, I assume that $\sigma^I \in [\sigma^l, \sigma^m]$.[21] Note that the delegation game includes uncertainty about I's choice, which may also impact salience through the aforementioned moral uncertainty. Such an effect might be implicitly in salience terms here, but that does not detract from delegation per se, especially in light of evidence that delegation affects behavior in the manner predicted by moral salience even in the absence of uncertainty (e.g., see results in Study 1 of Coffman, 2011).

Theorem 6.3.1: Suppose altruism is second order in the delegation game. Then, the fairest dictators choose to allocate fairly themselves, less fair ones delegate, and the least fair allocate unfairly themselves. Fewer dictators choose to allocate fairly in the delegation game than in the standard dictator game. The fraction of intermediaries allocating fairly lies between the fraction allocating fairly and the sum of the fractions allocating fairly or delegating in the delegation game.

Proof: See Appendix.

These predictions are consistent with the patterns observed in the BF treatments discussed above: some fair Ds in the baseline switch to delegation, when that is an option. The fraction of intermediaries, who allocate fairly is 18%, which lies in the predicted interval of the 17% of Ds allocating fairly and the 34% allocating fairly or delegating in the delegation game.

Most experimental studies of delegation also include the possibility of subsequent punishment of dictator decisions. Some of these results are summarized in the following SF.

SF 6.3.2: There is no significant punishment of fair choices, regardless of delegation. Dictators are punished significantly less for unfair allocations that result from delegation than from their own decisions, but delegation increases punishment of intermediaries for unfair choices (Coffman, 2011, Bartling and Fischbacher, 2012, Oexl and Grossman, 2013).

[21] Another question is whether to treat I as an agent or a patient. In the latter, but not the former, case, I's allocations must be included in the moral preferences of the D. In the baseline, it seems clear that I is wholly passive and, therefore, a patient. The same is true when the D chooses directly in the delegation game. For simplicity and consistency, therefore, and because the results do not depend qualitatively on this call, I is treated everywhere as a patient.

The following theorem considers costless punishment in the delegation game assuming non-strategic D choices as with unanticipated spectator punishment.

Theorem 6.3.2: Suppose the mean threshold for character is below the value for allocating fairly and above the level for allocating unfairly in the delegation game, inclusive. Then spectators do not punish those, who make no decisions or who directly choose fair outcomes. They do punish those, who directly choose unfair outcomes, whether in the delegation or the dictator game. In addition, if the threshold is above the expected character of a D who delegates, spectators will punish dictators who delegate, but less strongly than they punish unfair choices and even less strongly, if fair obtains after delegation.

Proof: See Appendix.

These predictions are summarized below, whereby z_P^{GCO} denotes the ideal punishment by the spectator, $\tilde{z}$, of person P in game G, who chose C resulting in, if uncertain, outcome O.

$$0 = z_D^{BF} = z_I^{BF} = z_I^{BU} > z_D^{BU}$$

$$0 = z_D^{DF} = z_I^{DF} = z_I^{DU} > z_D^{DU}$$

$$0 = z_I^{DDF} \geq z_D^{DDF} \geq z_D^{DDU} > z_I^{DDU}$$

The assumption that the character threshold lies below the level for allocating fairly and above the level for allocation unfairly in the delegation game is consistent with intuition and the condition that there be some degree of both reward and punishment. The further question of whether it is located sufficiently above the level for unfair choices that spectators punish delegation is an empirical matter.

I am unaware of any delegation experiment with unannounced spectator punishment, but in some treatments of BF, a randomly chosen R could spend one point to assign up to seven punishment points to each of the other three subjects for each possible decision (i.e., using the strategy method). In these treatments, punishment was common knowledge, which potentially confounds inferences about the motives behind D choices because of possible strategic self-interest. Nevertheless, if R estimates of D types in these treatments produces the same ranking of Ds as that of unannounced spectators, then the results remain qualitatively the same. Thus, this comparison must be taken with a grain of salt, but these results are worth examining given the negligible cost of punishment in BF, the absence of a compelling reason to believe the possibility of strategic self-interest would undo rankings of Ds, and the consistency of predictions with the

patterns observed in this study. The results for punishment in the BF experiment are summarized in Table 7 and are in accord with all of the predictions of Theorem 6.3.2. On the two theoretically ambiguous effects, DDF/D and DDU/D, there is a hint that some sanctioners wish to punish Ds for delegating, but the evidence on this is weak.

TABLE 7
PUNISHMENT POINTS IN BARTLING AND FISCHBACHER (2012)

Dictator Game:	BF/D	BF/I	BU/I	BU/D
Baseline	-0.41	-0.34	-0.42	-3.70
Delegation Game:	DF/D	DF/I	DU/I	DU/D
Allocate	-0.19	-0.15	-0.75	-4.27
Delegation Game:	DDF/I	DDF/D	DDU/D	DDU/I
Delegate	-0.20	-0.24	-1.31	-3.96

7. Moral Point Salience

So far in this paper, moral salience has referred to what is more properly called *moral set salience*, but now we turn briefly to what I will call *moral point salience*. As I use the terms here, set salience relates to properties of disjoint subsets of the decision context, viz., moral and non-moral context, whereas point salience refers to individual elements of the context. The latter provides a simple explanation for the well-established pattern from economics experiments of atoms at certain points in choice distributions. I will discuss three examples of actions that are chosen with greater frequency than alternative choices in their neighborhood. I believe these are important examples of point salience, but I do not claim that this is necessarily an exhaustive list. First, equal splits are a frequent choice in many experiments, such as in ultimatum and dictator games (Camerer, 2003). Second, zero transfers have also emerged as a frequent choice in dictator experiments where taking options rule out a corner solution as an explanation, e.g., List (2007), Cappelen et al. (2013), and Alevy et al. (2014). Finally, many studies have found that, when certain actions are explicitly highlighted (e.g., actions of previous subjects, experimenter suggestions, actions of role models), they tend to be chosen more frequently, including in dictator games, e.g., Andreoni and Bernheim (2009), public good games, e.g., Croson and Marks

(2001), COVID-19-related contributions and volunteering, e.g., Abel and Brown (2020), and field experiments on charitable contributions, e.g., Shang and Croson (2009).

Researchers have explained or modeled these patterns in various ways, but each account has its limitations, and I am unaware of a unified explanation for all three. For example, theoretical explanations for equal splits include a kinked inequality aversion term (Fehr and Schmidt, 1999), infinite inequity aversion on the part of some subjects (e.g., Konow, 2000), or a signaling game in which agents value social image (Andreoni and Bernheim, 2009). A strength of the first approach is that it provides a simple explanation for results from numerous bargaining and market experiments, but it is inconsistent with other findings, such as the frequent choice of transfers between zero and one-half in dictator experiments. The second approach accounts for the heterogeneity of types according to their degree of inequity aversion observed in many experiments and also accommodates concepts of equity other than equality, but it is inconsistent with variation in the percentage of subjects making equitable choices, for example, with the price of giving (Andreoni and Miller, 2002). The third approach offers a persuasive account for equal splits in the dictator game that avoids seemingly ad hoc assumptions about non-differentiability of utility, but it relies on a complicated theoretical apparatus the extension of which to other games is not straightforward. Moreover, none of these approaches provides explanations for all three examples above. Zero transfers, even in dictator games with taking, might be explained by dictators, who experience an endowment effect, but that explanation does nothing to account for the first or third examples. The masses at highlighted choices can be understood as focal points that facilitate coordination in strategic games, but that does not explain masses at dominated choices in non-strategic decisions. One might invoke experimenter demand effects (Zizzo, 2010), but they also do not provide a coherent and unified explanation for all three effects.

Moral point salience offers a simple and parsimonious account for these three types of masses based on moral considerations. Specifically, it concerns elements, P, of the set of actions, X, that are, for moral reasons, more salient than other elements of X, whereby $P \subset X$. Moral point salience is a term that is applied to fairness preferences, f, of agent i and takes the form:

$$s_i(x) = \begin{cases} \overline{s}_i \geq 1 \\ 1 \end{cases} \text{if} \begin{matrix} x \in P \\ x \notin P \end{matrix}.$$

That is, morally salient actions may be more heavily weighted in some agents' fairness preferences than other elements of the set of actions, whereby the weight can vary by person.

This discontinuity in utility can prove irksome, and, in fact, the analysis in this paper thus

far has not required point salience to be invoked, largely for that reason. Without point salience, the theory has been applied to explain a wide range of classic and anomalous results while retaining differentiability of the utility function. Nevertheless, point salience is helpful to account for masses that often materialize when examining the distribution of choices. So, I wish to outline and justify briefly the position that moral point salience not only earns first place in an Occam's razor contest to explain such masses but also that it is also a persuasive part of a coherent morals-based framework. I propose three categories of moral point salience below.

First, *norm salience*, where $\eta \in P$, is the most intuitive type of moral point salience. When first hearing about the standard dictator or ultimatum game, I suspect almost everyone thinks the same thing: the morally right choice is to split the stakes equally. We can torture ourselves for alternative, and more elegant, explanations, but I believe the most compelling one is staring us in the face: the morally preferred choice is obvious, in this case. I take obvious to imply, formally, that a discrete positive increment is applied to moral preferences of many agents for that choice. Of course, stakeholders, such as dictators or proposers, might make another choice due to self-interest, but, as stated in Assumption 3, the moral norm, η, can be identified from the choices of spectators. A concrete and intuitive way to operationalize this is to associate the entitlement in experiments with the modal choice of spectators and the salience of the entitlement as being in direct proportion to spectator consensus, specifically, whereby consensus can be conceptualized as inversely related to variance in spectator judgments (Konow, 2009).

As discussed in section 3.1, the norm defaults to equality in simple decision contexts, like the standard dictator game. In fact, equal splits emerge frequently in most of the games examined in this paper. But what if the norm is not as simple and obvious as equality? As argued in section 3.1, when the context provides information relevant to other norms, behavior shifts towards those norms, such as efficiency (see section 4.2), but does that produce masses at those norms? Consider a more demanding test of point salience based on a more complicated rule: equity calls for allocations that are proportional to contributions. Figure 12 summarizes results from experiments that illustrate point salience. In Konow, Saijo and Akai (2020) subjects first perform a real effort task, and then dictators allocate the resulting earnings. Panel a shows the amounts stakeholding dictators allocate to recipients, and panel b shows the amounts spectators allocate to one member of each pair. Specifically, the horizontal axis represents the difference in points allocated to recipients from the amount they produced, which is their entitlement and the

equitable amount. The mode and choice of 50% of stakeholders in panel a and of 52% of spectators in panel b is the equitable amount. Thus, in general, with norm salience, P is a rule and x a point that might be conditioned on another variable in the context. For example, when the norm is equity, x is conditioned on individual contributions, and when the norm is basic needs, it is conditioned on individual needs. Konow (2001) and Konow, Saijo and Akai (2020) argue that equality is the norm by default, when there is no or insufficient information to apply another norm.

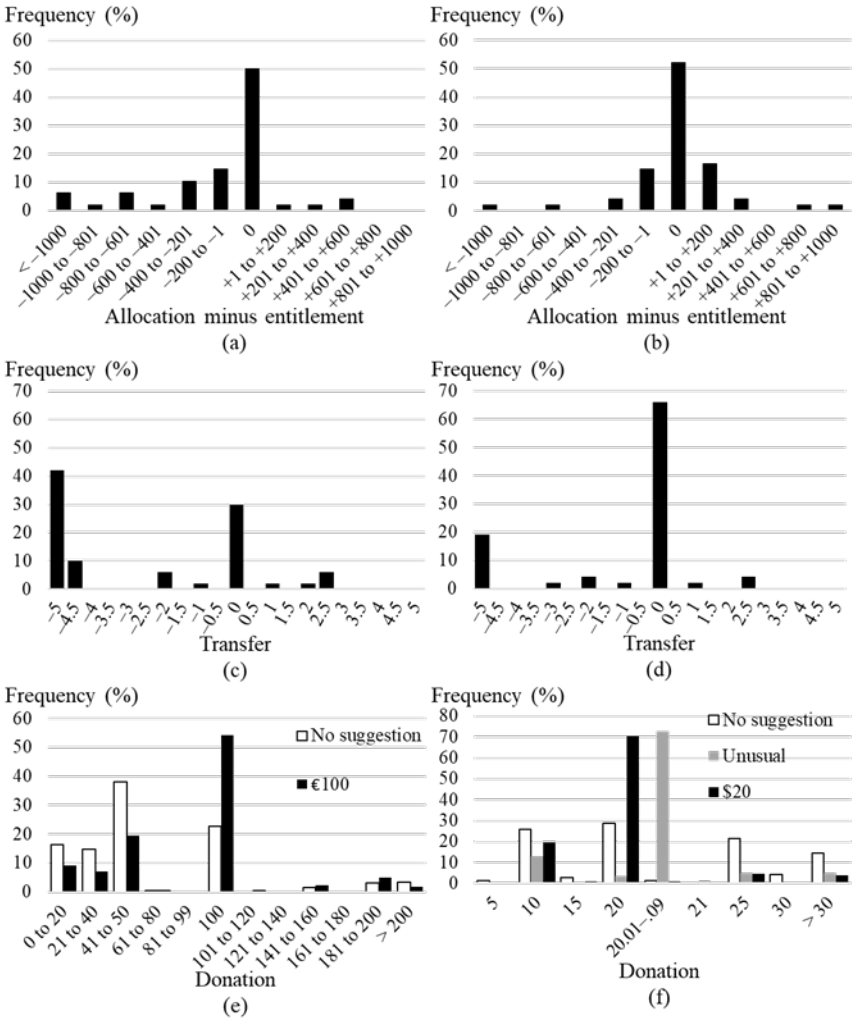

FIGURE 12. – Moral point salience.

Sources: Konow. Saijo and Akai (2020) stakeholders (a) and spectators (b), List (2007) Take $5 (c) and Earnings (d), Adena, Huck and Rasul (2014) No suggestion and €100 (e), and Edwards and List (2014) No ask, $20 ask and Unusual ask (f).

Second, *null salience* is the salience of inaction, i.e., the choice neither to help nor to harm, such as neither giving nor taking in a dictator game, denoted $0 \in P$. Null salience is related to the distinction in ethics between sins of commission for the wrongs one chooses versus

sins of omission for the acts one should perform but does not. Various experiments in economics and psychology suggest people have a stronger aversion to acts of commission than to ones of omission (e.g., Cox, Servátka, and Vadovič, 2017, Spranca, Minsk, and Baron, 1991). That is, an individual, who otherwise might prefer to harm another, say, take $1 in a dictator game, might experience a discontinuity in the marginal moral disutility of doing so. One way to model this is with moral point salience, where utility is discretely greater at zero. Of course, in many experiments, such as the standard dictator game, a mass at zero logically emerges as a corner solution due to selfish, spiteful or insufficiently fair dictators. But such censoring is not a problem in dictator games with taking, and, in fact, those studies typically find a mass at zero. This is illustrated in panels c and d of Figure 12, which depicts the Take $5 and Earnings treatments, respectively, of List (2007), and show between 30% and 66% of dictators choosing inaction, i.e., transfers of zero.[22]

Third, and finally, *threshold salience* refers to the action, $\tilde{x} \in P$, that corresponds to the agent's preferred character threshold, $\tilde{\gamma}$, given the context and its moral set salience. Remember that $\tilde{x}$ is the action that is neither praiseworthy nor blameworthy and that it is less than the objectively fair transfer, η, if salience is below a sufficiently high level. Thus, this can be thought of as the action that is "fair enough" given the context. In the case of a stakeholder, we can think of this as the stakeholder's (potentially biased) belief about what constitutes "fair enough." The current theory assumes heterogeneity in the value of $\tilde{\gamma}$, even among spectators, so it is not clear why a mass would materialize at any particular stakeholder choice. But experimental evidence establishes that beliefs about the sufficient level of norm compliance, $\tilde{x}$, are malleable and can be manipulated, e.g., Bicchieri and Chavez (2010) and Bicchieri, Dimant, Gächter and Nosenzo (2020). Indeed, compliance with norms responds to and sometimes coalesces around information, including about past trusting behavior of others (Berg, Dickhaut and McCabe, 1995), recommended contributions to a public good (e.g., Croson and Marks, 2001) and default levels of transfers to recipients in dictator games (e.g., Andreoni and Bernheim, 2009). Such evidence is consistent with an effect of specific information on beliefs about appropriate norm compliance and, as a result, on behavior itself.

[22] Another way this could be modeled is as a kink in the altruism function at zero. This is consistent with a mass at zero but one disadvantage of this approach is that a kink is not consistent with the paucity of transfers typically observed just above and just below zero whereas point salience is.

Voluntary contributions to charities or public goods lend themselves to examination of this effect, since many people feel morally obliged to donate but plausibly have non-degenerate distributions of beliefs about the appropriate amount. Studies of such contributions that include suggested donations yield evidence consistent with threshold salience. Panel e of Figure 12 shows contributions to a public good (viz., an opera house) in a field experiment of Adena, Huck and Rasul (2014). When solicitations explicitly suggest a €100 contribution, the fraction of such donations is significantly greater than that when no suggestion is made ($p<0.001$), according to the two-tail z-tests used in all comparisons here.[23] The results of the field experiment of Edwards and List (2014) on alumni donations to a university are summarized in panel f of Figure 12. The fraction of $20 contributions is significantly greater ($p<0.001$), when that amount is explicitly stated in solicitations. A further treatment in which solicitations suggest unusual amounts, like $20.01 or $20.04, produce a similar increase in the frequency of choices of stated amounts ($p<0.001$), corroborating the robustness of this effect, even when suggestions are not round numbers.

An advantage of the way these three types of moral point salience are formulated is that they can be specified and identified empirically. Norm salience can be inferred from spectator choices, null salience defines itself, and threshold salience can be inferred from behavioral responses to information or from incentivized solicitation of beliefs. This discussion on moral point salience indicates some commonsensical concepts to account for certain patterns that are not explained by moral set salience and indicates possible avenues for future research.

8. Conclusions

This paper proposes a tractable theory to explain not only classic results on allocative preferences and reciprocity but also a wide range of anomalous findings about moral behavior, including moral proximity, moral uncertainty, the outcome bias, the taking effect, joy of destruction, moral egress, willful ignorance and delegation. At various stages, I have discussed alternative explanations for specific phenomena, such as experimental artefacts (e.g., Bardsley, 2008), motivated reasoning (e.g., Gino et al., 2016) and image concerns (e.g., Andreoni and

[23] A further treatment shows that, when solicitations suggest €200, the effect dissipates. This seems consistent with agents having prior beliefs about the distribution of appropriate donations, whereby suggestions provide signals that impact $\overline{s}_i$ in direct relationship to their proximity to priors such that outlier suggestions are ineffectual. Nevertheless, formal analysis of this question goes beyond the scope of this paper.

Bernheim, 2009), including what I see as the strengths of those alternatives. As stated at the start, the goal is not to dismiss or conduct a beauty contest with other accounts of specific phenomena. Instead, one goal was to present an until now neglected explanation, which plausibly sweeps up much of the variance in observed behavior. Another goal was to illustrate the theory's flexibility and ease of application, that is, to argue its appeal on the basis of Occam's razor. A related aim was to demonstrate the generality of the theory across an unprecedented set of sometimes enigmatic empirical results on moral preferences. Finally, the theory was tested out-of-sample and its predictions corroborated in a new experiment.

Future research could explore possible roles for moral salience and virtue preferences in relation to other types of moral preferences apart from allocative preferences, e.g., trust, trustworthiness, honesty, and cooperation. Further work could also analyze the factors that affect how different moral and non-moral contexts might be integrated across different decisions at a point in time as well as over time. That is, one could examine the effects on moral salience of presenting similar decisions while varying the moral and non-moral context, which could, for example, account for order effects. In addition, this paper focused on non-strategic decision-making in order to simplify the analysis and avoid factors that might confound inferences about the forces being studied. But future work might extend the theory to situations involving strategic interaction, such as bargaining. A theory incorporating moral salience and virtue preferences could be applied to decision-making in experimental games, like the ultimatum game, trust game, moonlighting game, centipede game, and public good games.

REFERENCES

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner (2000): "The Moonlighting Game: An Experimental Study on Reciprocity and Redistribution," *Journal of Economic Behavior and Organization,* 42(2), 265-277.

Abbink, Klaus and Abdolkarim Sadrieh (2009): "The Pleasure of Being Nasty," *Economic Letters,* 105(3), 306-308.

Abbink, Klaus and Benedikt Herrmann (2009): "Pointless Vendettas," *SSRN Electronic Journal,* 1-11.

Abbink, Klaus and Benedikt Herrmann (2011): "The Moral Costs of Nastiness," *Economic Inquiry,* 49(2), 631-633.

Abel, Martin and Willa Brown (2020): "Prosocial Behavior in the Time of COVID-19: The Effect of Private and Public Role Models," *IZ Discussion Paper,* 13207, 1-26.

Adena, Maja, Steffen Huck, and Imran Rasul (2014): "Charitable Giving and Nonbinding Contribution-Level Suggestions — Evidence from a Field Experiment," *Review of Behavioral Economics,* 1(3), 275-293.

Aguiar, Fernando, Alice Becker, and Luis Miller (2013): "Whose Impartiality? An Experimental Study of Veiled Stakeholder, Involved Spectators and Detached Observers," *Economics and Philosophy,* 29 (2), 155-174.

Alevy, Jonathan E., Francis L. Jeffries, and Yonggang Lu (2014): "Gender – and Frame-Specific Audience Effects in Dictator Games," *Economic Letters,* 122, 50-54.

Almås, Ingvild, Alexander Cappelen, and Bertil Tungodden (2020): "Cutthroat Capitalism versus Cuddly Socialism: Are American More Meritocratic and Efficiency-seeking than Scandinavians?." *Journal of Political Economy,* 128 (5), 1753-1788.

Almenberg, Johan, Anna Dreber, Coren L. Apicella, and David G. Rand (2011): "Third Party Reward and Punishment: Group Size, Efficiency and Public Goods," *Psychology of Punishment,* 10, 1-17.

Andreoni, James (1989): "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy,* 97(6), 1447-1458.

Andreoni, James and B. Douglas Bernheim (2009): "Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica,* 77(5), 1607-1636.

Andreoni, James and John Miller (2002): "Giving According to Garp: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica,* 70(2), 737-753.

Andreoni, James, Justin M. Rao, and Hannah Trachtman (2017): "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving," *Journal of Political Economy,* 125(3), 625-653.

Arechar, Antonio A., Simon Gächter, and Lucas Molleman (2018): "Conducting Interactive Experiments Online," *Experimental Economics,* 21, 99-131.

Ashraf, Nava, and Oriana Bandiera (2017): "Altruistic Capital," *American Economic Review: Papers & Proceedings,* 107(5), 70-75.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer (1995): "Biased Judgements of Fairness in Bargaining," *The American Economic Review,* 85(5), 1337-1343.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2010): "Social Incentives in the Workplace," *The Review of Economic Studies,* 77, 417-458.

Bardsley, Nicholas (2008): "Dictator Game Giving: Altruism or Artefact?," *Experimental Economics,* 11(2), 122-133.

Bartling, Björn and Urs Fischbacher (2012): "Shifting the Blame: One Delegation and Responsibility," *Review of Economic Studies,* 79, 67-87.

Bartling, Björn, Florian Engl, and Roberto A. Weber (2014): "Does Willful Ignorance Deflect Punishment? - An Experimental Study," *European Economic Review,* 70, 512-524.

Bénabou, Roland and Jean Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review,* 96(5), 1652-1678.

Benz, Mathias and Stephan Meier (2008): "Do People Behave in Experiments as in the Field? – Evidence from Donations," *Experimental Economics,* 11(3), 268-281.
Berg, Joyce, John Dickhaut, and Kevin McCabe (1995): "Trust, Reciprocity, and Social History," *Games and Economic Behavior,* 10, 122-142.
Bertrand, Marianne and Sendhil Mullainathan (2001): "Are CEOs Rewarded for Luck? The Ones Without Principals Are," *Quarterly Journal of Economics,* 116(3), 901-932.
Bicchieri, Cristina and Alex Chaves (2010): "Behaving as Expected: Public Information and Fairness Norms," *Journal of Behavioral Decision Making,* 23(2), 161-178.
Bicchieri, Cristina, Eugen Dimant, Simon Gächter, and Daniele Nosenzo (2020): "Observability, Social Proximity, and the Erosion of Norm Compliance," *CESifo Working Paper,* 8212, 1-32.
Bohnet, Iris and Bruno S. Frey (1999): "Social Distance and Other-Regarding Behavior in Dictator Games: Comment," *The American Economic Review,* 89, 335-339.
Bolton, Gary E., Jordi Brandts, and Axel Ockenfels (2005): "Fair Procedures: Evidence from Games Involving Lotteries," *The Economic Journal,* 115(506), 1054-1076.
Bontol, Gary E., and Elena Katok (1998): "An Experimental Test of the Crowding Out Hypothesis: The Nature of Beneficent," *Journal of Economic Behavior & Organization,* 37(3), 315-331.
Bolle, Friedel, Johnathan H.W. Tan, and Daniel John Zizzo (2014): "Vendettas," *American Economic Journal: Microeconomics,* 6(2), 93-130.
Bordalo, Pedro, Nicola Gennaioli, and Andrei Sheifer (2012): "Salience in Experimental Tests of the Endowment Effect," *American Economic Review,* 102(3), 47-52.
Bordalo, Pedro, Nicola Gennaioli, and Andrei Sheifer (2013): "Salience and Consumer Choice," *Journal of Political Economics,* 121(5), 803-843.
Bordalo, Pedro, Nicola Gennaioli, and Andrei Sheifer (2016): "Competition for Attention," *Review of Economic Studies,* 83(2), 481-513.
Brañas-Garza, Pablo (2007): "Promoting Helping Behavior with Framing in Dictator Games," *Journal of Economic Psychology,* 28, 477-486.
Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson (2007): "Is Generosity Involuntary?," *Economic Letters,* 94, 32-37.
Brock, J. Michelle, Andreas Lange, and Erkut Y. Ozbay (2013): "Dictating the Risk: Experimental Evidence on Giving in Risky Environments," *American Economic Review,* 103, 415-437.
Brownback, Andy and Michael A. Kuhn (2019): "Understanding Outcome Bias," *Games and Economic Behavior,* 117, 342-360.
Candelo, Natalie, Catherine Eckel, and Cathleen Johnson (2018): "Social Distance Matters in Dictator Games: Evidence from 11 Mexican Villages," *Games,* 9(77), 1-13.
Camerer, Colin (2003): *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton: Princeton University Press.
Camerer, Colin, and Richard H. Thaler (1995): "Anomalies: Ultimatums, Dictators and Manners," *Journal of Economic Perspectives,* 9(2), 209-219.
Campos-Mercade, Pol, Armando N. Meier, Florian H. Schneider, and Erik Wengström (2020): "Prosociality Predicts Health Behaviors During the COVID-19 Pandemic," Department of Economics Working Paper No. 346, University of Zurich.
Cappelen, Alexander W., James Konow, Erik O. Sorensen, and Bertil Tungodden (2013): "Just Luck: An Experimental Study of Risk-Taking and Fairness," *American Economic Review,* 103(4), 1398-1413.
Charness, Gary and David I. Levine (2007): "Intention and Stochastic Outcomes: An Experimental Study," *The Economic Journal,* 117(522), 1051-1072.
Charness, Gary and Martin Dufwenberg (2006): "Promises and Partnership," *Econometrica,* 74(6), 1579-1601.
Charness, Gary and Matthew Rabin (2002): "Understanding Social Preferences with Simple Tests," *The Quarterly Journal of Economics,* 117(3), 817-869.
Charness, Gary and Uri Gneezy (2008): "What's in a Name? Anonymity and Social Distance in Dictator and Ultimatum Games," *Journal of Economic Behavior and Organization,* 68, 29-35.

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016): "oTree – An Open-source Platform for Laboratory, Online, and Field Experiments," *Journal of Behavioral and Experimental Finance,* 9, 88-97.
Chetty, Raj, Adam Looney, and Kory Kroft (2009): "Salience and Taxation: Theory and Evidence," *American Economic Review,* 99(4), 1145-1177.
Cherry, Todd L., Peter Frykblom, and Jason F. Shogren (2002): "Hardnose the Dictator," *The American Economic Review,* 92(4), 1218-1221.
Chowdhury, Subhasish M., Joo Young Jeon, and Bibhas Saha (2017): "Gender Differences in the Giving and Taking of the Dictator Game," *Southern Economic Journal,* 84(2), 474-483.
Coffman, Lucas C. (2011): "Intermediation Reduces Punishment (and Reward)," *American Economic Journal: Microeconomics,* 3(4), 77-106.
Cox, James C. (2004): "How to Identify Trust and Reciprocity," *Games and Economic Behavior,* 46(2), 260-281.
Cox, James C., Maroš Servátka, and Radovan Vadovič (2017): "Status Quo Effects in Fairness Games: Reciprocal Responses to Acts of Commission Versus Acts of Omission," *Experimental Economics,* 20, 1-18.
Cox, James C., John A. List, Michael Price, Vjollca Sadiraj, and Anya Samek (2019): "Moral Costs and Rational Choice: Theory and Experimental Evidence," *Experimental Economics Center Working Paper Series,* 2, 1-52.
Crawford, Vincent P., and Nagore Iriberri (2007), "Fatal Attraction: Salience, Naïveté, and Sophistication in Experimental "Hide-and-Seek" Games," *American Economic Review,* 97(5), 1731-1750.
Crawford, Vincent P., Uri Gneezy, and Yuval Rottenstreich (2008): "The Power of Focial Points is Limited: Even Minute Payoff Asymmetry May Yield Coordination Failures," *American Economic Review,* 98(4), 1443-1448.
Croson, Rachel, and James Konow (2009): "Social Preferences and Moral Biases," *Journal of Economic Behavior and Organization*, 69(3), 201-212.
Croson, Rachel and Melanie Marks (2001): "The Effect of Recommend Contributions in the Voluntary Provision of Public Goods," *Economic Inquiry,* 39(2), 238-249.
Crumpler, Heidi and Philip J. Grossman (2008): "An Experimental Test of Warm Glow Giving," *Journal of Public Economics,* 92(5-6), 1011-1021.
Cushman, Fiery, Anna Dreber, Ying Wang, and Jay Costa (2009): "Accidental Outcomes Guide Punishment in a 'Trembling Hand' Game," *PLoS ONE,* 4(8), 1-7.
Dahlgren, Peter M., Adam Shehata, and Jesper Strömbäck (2019): "Reinforcing Spirals at Work? Mutual Influences between Selective New Exposure and Ideological Leaning," *European Journal of Communications,* 34(2), 159-174.
Dal Bó, Ernesto, and Pedro Dal Bó (2014): "Do the Right Thing: The Effects of Moral Suasion on Cooperation," *Journal of Public Economics,* 117, 28-38.
Dal Bó, Pedro, and Guillaume R. Fréchette (2018): "On the Determinants of Cooperation in Infinitely Repeated Games: A Survey," *Journal of Economic Literature,* 56(1), 60-114.
Dana, Jason, Daylian M. Cain, and Robyn M. Dawes (2006): "What You Don't Know Won't Hurt Me: Costly (But Quiet) Exit in Dictator Games," *Organizational Behavior and Human Decision Processes,* 100(2), 193-201.
Dana, Jasion, Roberto A. Weber, Jason Xi Kuang (2007): "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory,* 33, 67-80.
Dejean, Sylvain (2020): "The Role of Distance and Social Networks in the Geography of Crowfunding: Evidence from France," *Regional Studies,* 54(3) 329-339.
DellaVigna, Stefano, John A. List, and Ulrike Malmendier (2012): "Testing for Altruism and Social Pressure in Charitable Giving," *The Quarterly Journal of Economics,* 127, 1-56.
De Oliveria, Angela C.M., Alexander Smith, and John Spraggon (2017): "Reward the Lucky? An Experimental Investigation of the Impact of Agency and Luck on Bonuses," *Journal of Economic Psychology,* 62, 87-97.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth (2018): "Measuring and Bounding Experimenter Demand," *American Economic Review,* 108(11), 3266-3302.
Dreber, Anna, Tore Ellingsen, Magnus Johannesson, and David G. Rand (2013): "Do People Care About Social Context? Framing Effects in Dictator Games," *Experimental Economics,* 16(3), 349-371.
Dufwenberg, Martin and Georg Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior,* 47(2), 268-298.
Eckel, Catherine C. and Philip J. Grossman (1996): "The Relative Price of Fairness: Gender Differences in a Punishment Game," *Journal of Economic Behavior and Organization,* 30(2), 143-158.
Edwards, James T. and John A. List (2014): "Toward an Understanding of Why Suggestions Work in Charitable Fundraising: Theory and Evidence from a Natural Field Experiment," *Journal of Public Economics,* 114, 1-13.
Ellingsen, Tore, Magnus Johannesson, Sigve Tjotta, Gaute Torsvik (2010): "Testing Guilt Aversion," *Games and Economic Behavior,* 68, 95-107.
Ellingsen, Tore and Magnus Johannesson (2008): "Anticipated Verbal Feedback Induces Altruistic Behavior," *Evolution and Human Behavior,* 29(2), 100-105.
Engel, Christoph (2011): "Dictator Games: A Meta Study," *Experimental Economics,* 14(4), 583-610.
Engelmann, Dirk and Martin Strobel (2004): "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments," *The American Economic Review,* 94(4), 857-869.
Falk, Armin and Urs Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior,* 54(2), 293-315.
Faravelli, Marco (2007): "How Context Matters: A Survey Based Experiment on Distributive Justice," *Journal of Public Economics,* 91(7-8), 1399-1422.
Fehr, Dietmar (2018): "Is Increasing Inequality Harmful? Experimental Evidence," *Games and Economic Behavior,* 107, 123-134.
Fehr, Ernst, Georg Kirchsteiger and Arno Riedl (1993): "Does Fairness Prevent Market Clearing? An Experimental Investigation," *The Quarterly Journal of Economics,* 108(2), 437-459.
Fehr, Ernest and Klaus M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics,* 114(3), 817-868.
Fehr, Ernest and Urs Fischbacher (2004): "Third-Party Punishment and Social Norms," *Evolution and Human Behavior,* 25(2), 63-87.
Feiler, Lauren (2014): "Testing Models of Information Avoidance with Binary Choice Dictator Games," *Journal of Economic Psychology,* 45, 253-267.
Finus, Michael and Pedro Pintassilgo (2013): "The Role of Uncertainty and Learning for the Success of International Climate Agreements," *Journal of Public Economics,* 103, 29-43.
Forsythe, Robert, Joel L. Horowitz, N. E. Savin, and Martin Sefton (1994): "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior,* 6(3), 347-369.
Franzen, Axel, and Sonja Pointner (2013): "The External Validity of Giving in the Dictatorship Game: A Field Experiment Using the Misdirected Letter Technique," *Experimental Economics,* 16(2), 155-159.
Gino, Francesca, Lisa L. Shu, and Max H. Bazerman (2010): "Nameless + Harmless = Blameless: When Seemingly Irrelevant Factors Influence Judgement of (Un)Ethical Behavior," *Organizational Behavior and Human Decision Processes,* 111(2), 93-101.
Gino, Francesca, Michael I. Norton. And Roberto A. Weber (2016): "Motivated Bayesians: Feeling Moral While Acting Egoistically," *Journal of Economic Perspectives,* 30(3), 189-212.
Golman, Russell, David Hagmann, and George Loewenstein (2017): "Information Avoidance," *Journal of Economic Literature,* 55, 96-135.
Green, Stuart P. (2007): "Looting, Law, and Lawlessness," *Tulane Law Review,* 81, 1129-1179.
Grossman, Philip J. and Catherine C. Eckel (2015): "Giving Versus Taking for Cause," *Economic Letters,* 132(C), 28-30.
Grossman, Zachary (2014): "Strategic Ignorance and the Robustness of Social Preferences," *Management Science,* 60(11), 2659-2665.

Grossman, Zachary (2015): "Self-Signaling and Social-Signaling in Giving," *Journal of Economic Behavior and Organization,* 117, 26-39.
Grossman, Zachary and Joël J. van der Weele (2017): "Self-Image and Willful Ignorance in Social Decisions," *Journal of the European Economic Association,* 15, 173-217.
Gurdal, Mehmet Y., Joshua B. Miller, and Aldo Rustichini (2013): "Why Blame?," *Journal of Political Economy,* 121(6), 1205-1247.
Güth, Werner, Steffen Huck, and Wieland Müller (2001): "The Relevance of Equal Splits in Ultimatum Games," *Games and Economic Behavior,* 37, 161-169.
Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982): "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization,* 3(4), 367-388.
Hammann, John R., George Loewenstein, and Roberto A. Weber (2010): "Self-Interest Through Delegation: An Additional Rationale for the Principal-Agent Relationship," *American Economic Review,* 100(4), 1826-1846.
Healy Andrew J., Neil Malhorta, and Cecilia Hyunjung Mo (2010): "Irrelevant Events Affect Voters' Evaluations of Government Performance," *Proceedings of the National Academy of Sciences of the United States of America,* 107(29), 12804-12809.
Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith (1996): "Social Distance and Other-Regarding Behavior in Dictator Games," *The American Economic Review,* 86(3), 653-660.
Horton, John J., David G. Rand, and Richard J. Zeckhauser (2011): "The Online Laboratory: Conducting Experiments in a Real Labor Market," *Experimental Economics,* 14(3), 399-425.
Iriberri, Nagore and Pedro Rey-Biel (2013): "Elicited Beliefs and Social Information in Modified Dictator Games: What do Dictators Believe Other Dictators do?," *Quantitative Economics,* 4(3), 515-547.
Kausel, Edgar E., Santiago Ventura, and Arturo Rodríguez (2019): "Outcome Bias in Subjective Ratings of Performance: Evidence from the (Football) Field," *Journal of Economic Psychology,* 75(B), 1-9.
Kessler, Esther, Maria Ruiz-Martos, and David Skuse (2012): "Destructor Game," *Working Papers,* 11, 1-9.
Khazan, Olga (2020): "Why People Loot," *The Atlantic,* June 2, 2020.
Kimbrough, Erik O. and Alexander Voxtroknutov (2016): "Norms Make Preferences Social," *Journal of the European Economic Association,* 14(3), 608-638.
Konow, James (2000): "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions," *American Economic Review*, 90(4), 1072-1091.
Konow, James (2001): "Fair and Square: The Four Sides of Distributive Justice," *Journal of Economic Behavior and Organizations*, 46(2), 137-164.
Konow, James (2005): "Blind Spots: The Effects of Information and Stakes on Fairness Bias and Dispersion," *Social Justice Research,* 18(4), 349-390.
Konow, James (2009): "Is Fairness in the Eye of the Beholder? An Impartial Spectator Analysis of Justice," *Social Choice and Welfare,* 33, 101-127.
Konow, James (2010): "Mixed Feelings: Theories of and Evidence on Giving," *Journal of Public Economics,* 94(3-4), 279-297.
Konow, James (2012): "Adam Smith and the Modern Science of Ethics," *Economics and Philosophy,* 28(3), 333-362.
Konow, James (2019): "Can Ethics Instruction Make Economics Students More Pro-Social?," *Journal of Economic Behavior and Organization,* 166, 724-734.
Konow, James, Tatsuyoshi Saijo, and Kenju Akai (2020): "Equity Versus Equality: Spectators, Stakeholders and Groups," *Journal of Economic Psychology,* 77, 1-28.
Korenok, Oleg, Edward L. Millner, and Laura Razzolini (2012): "Are Dictators Averse to Inequality?," *Journal of Economic Behavior and Organization,* 82(2-3), 543-547.
Korenok, Oleg, Edward L. Millner, and Laura Razzolini (2014): "Taking, Giving, and Impure Altruism in Dictator Games," *Experimental Economics,* 17(3), 488-500.
Korenok, Oleg, Edward L. Millner, and Laura Razzolini (2017): "Feelings of Ownership in Dictator Games," *Journal of Economic Psychology,* 61, 145-151.

Korenok, Oleg, Edward L. Millner, and Laura Razzolini (2018): "Taking Aversion," *Journal of Economic Behavior and Organization,* 150, 397-403.
Krawczyk, Michal and Fabrice Le Lec (2010): "'Give Me a Chance!'An Experiment in Social Decision Under Risk," *Experimental Economics,* 13(4), 500-511.
Krupka, Erin L. and Roberto A. Weber (2013): "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?," *Journal of the European Economic Association,* 11(3), 495-524.
Kühl, Leonie and Nora Szech (2017): "Physical Distance and Cooperativeness Towards Strangers," *CESifo Working Paper,* 6825, 1-64.
Larson, Tara and C. Monica Capra (2009): "Exploiting Moral Wiggle Room: Illusory Preference for Fairness? A Comment," *Judgement and Decision Making,* 4(6), 467-474.
Lazear, Edward P., Urlike Malmendier, and Roberto A. Weber (2012): "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics,* 4, 136-163.
Levine, David K. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics,* 1, 593-622.
List, John A. (2007): "One the Interpretation of Giving in Dictator Games," *Journal of Political Economy,* 115(3), 482-493.
Mollerstrom, Johanna, Bjorn-Atle Reme and Erik O. Sorensen (2015): "Luck, Choice and Responsibility – An Experimental Study of Fairness Views," *Journal of Public Economics,* 131, 33-40.
Muller, Daniel and Sander Renes (2017): "Fairness Views and Political Preferences – Evidence from a Large *Online* Experiment," *Working Papers in Economics and Statistics,* 10, 1-39.
Oexl, Regine and Zachary J. Grossman (2013): "Shifting the Blame to a Powerless Intermediary," *Experimental Economics,* 16(3), 306-312.
Offerman, Theo (2002): "Hurting Hurts More Than Helping Helps," *European Economic Review,* 46(8), 1423-1437.
Oxoby, Robert J. and John Spraggon (2008): "Mine and Yours: Property Rights in Dictator Games," *Journal of Economic Behavior,* 65(3-4), 703-713.
Peterson, Erik, Sharad Goel, and Shanto Iyengar (2019): "Partisan Selective Exposure in Online News Consumption: Evidence from the 2016 Presidential Campaign," *Politcal Science Research and Methods,* 1-17.
Quarantelli, E. L. and Russell R. Dynes (1968): "Looting in Civil Disorders: An Index of Social Change," *The American Behavioral Scientist,* 11, 7-10.
Rabin, Matthew (1993): "Incorporating Fairness into Game Theory and Economics," *The American Economic Review,* 83(5), 1281-1302.
Rey-Biel, Pedro, Roman Sheremeta, and Neslihan Uler (2018): "When Income Depends on Performance and Luck: The Effects of Culture and Information on Giving," *Experimental Economics and Culture (Research in Experimental Economics, vol. 20),* Bingley, UK: Emerald Publishing Ltd, 167-203.
Rigdon, Mary, Keiko Ishii, Motoki Watabe, Shinobu Kitayama (2009): "Minimal Social Cues in the Dictator Game," *Journal of Economic Psychology,* 30(3), 358-367.
Rubin, Jared and Roman Sheremeta (2016): "Principal-Agent Settings with Random Shocks," *Management Science,* 62(4), 985-999.
Serra-Garcia, Marta and Nora Szech (2019): "The (In)Elasticity of Moral Ignorance," *KIT Working Paper Series in Economics,* 134, 1-60.
Sezer, Ovul, Ting Zhang, Francesca Gino, Max H. Bazerman (2016): "Overcoming the Outcome Bias: Making Intentions Matter," *Organizational Behavior and Human Decision Processes,* 137, 13-26.
Shang, Jen and Rachel Croson (2009): "A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods," *The Economic Journal,* 119(540), 1422-1439.
Smith, Alexander (2015): "On the Nature of Pessimism in Taking and Giving Games," *Journal of Behavioral and Experimental Economics,* 54, 50-57.

Sobel, Joel (2005): "Interdependent Preferences and Reciprocity," *Journal of Economic Literature,* 43, 392-436.
Spiekermann, Kai and Arne Weiss (2016): "Objective and Subjective Compliance: A Norm-Based Explanation of 'Moral Wiggle Room'," *Games and Economic Behavior,* 96, 170-183.
Spranca, Mark, Elisa Minsk, and Jonathan Baron (1991): "Omission and Commission in Judgment and Choice," *Journal of Experimental Social Psychology,* 27, 76-105.
Sutter, Matthias, Jürgen Huber, Michael Kirchner, Matthias Stefan, and Markus Walzl (2020): "Where to Look for the Morals in Markets," *Experimental Economics,* 23, 30-52.
Touré-Tillery, Maferima, and Ayelet Fishbach (2017): "Too Far to Help: The Effect of Perceived Distance on the Expected Impact and Likelihood of Charitable Action," *Journal of Personality and Social Psychology,* 112(6), 860-876.
Van Koten, Silvester, Andreas Ortmann, and Vitezslav Babicky (2013): "Fairness in Risky Environments: Theory and Evidence," *Games,* 4(2), 208-242.
Walzer, Michael (1983): "Spheres of Justice: A Defense of Pluralism and Equality," *The Journal of Philosophy,* 83(8), 457-468.
Whitt, Sam and Rick K. Wilson (2007): "The Dictator Game, Fairness and Ethnicity in Postwar Bosnia," *American Journal of Political Science,* 51(3), 655-668.
Williams, Bernard (1981): *Moral Luck*. Cambridge: Cambridge University Press.
Xiao, Erte and Daniel Houser (2009): "Avoiding the Sharp Tongue: Anticipated Written Messages Promote Fair Economic Exchange," *Journal of Economic Psychology,* 30(3) 393-404.
Zhang, Le, and Andreas Ortmann (2013): "On the Interpretation of Giving, Taking, and Destruction in Dictator Games and Joy-of-Destruction Games," Australian School of Business Research Paper No. 2012ECON50A (http://dx.doi.org/10.2139/ssrn.219040).
Zizzo, Daniel John and Andrew J. Oswald (2001): "Are People Willing to Pay to Reduce Others' Income," *Annales d'Économie et de Statistique,* 63, 39-65.
Zizzo, Daniel John (2003): "Money Burning and Rank Egalitarianism with Random Dictators," *Economic Letters,* 81(2), 263-266.
Zizzo, Daniel John (2010): "Experimenter Demand Effects in Economic Experiments," *Experimental Economics,* 13, 75-98.

Appendix: Proofs

Proof of Theorem 2.2.1:

$$dU/dx = -u'(X-x) + \sigma\phi f'\big(\phi(Y+x-\eta)\big) + \sigma\alpha g'(\alpha x) = 0$$

Applying the implicit function theorem to solve for $x(\sigma)$, substituting into the first order condition, and differentiating with respect to σ,

$$u''\frac{dx}{d\sigma} + \phi f' + \sigma\phi^2 f''\frac{dx}{d\sigma} + \alpha g' + \sigma\alpha^2 g''\frac{dx}{d\sigma} = 0$$

$$dx/d\sigma = \frac{-\phi f' - \alpha g'}{u'' + \sigma_f \phi^2 f'' + \sigma_g \alpha^2 g''} > 0.$$

Proof of Theorem 2.2.2

By Theorem 2.2.1, $\frac{dx}{d\sigma} > 0$ and we assume $\frac{d^2x}{d\sigma^2} \geq 0$. Write $x(\sigma)$ by the implicit function theorem and note that, by assumption, $\sigma(p,n)$, $\frac{\partial\sigma}{\partial n} < 0$ and $\frac{\partial^2\sigma}{\partial n^2} > 0$. Then, we can write the composite function $x\big(\sigma(p,n)\big)$, and, by the chain rule

$$\frac{\partial x}{\partial n} = \frac{dx}{d\sigma}\frac{\partial\sigma}{\partial n} < 0.$$

Taking the second derivative,

$$\frac{\partial^2 x}{\partial n^2} = \frac{d^2x}{d\sigma^2}\left(\frac{\partial\sigma}{\partial n}\right)^2 + \frac{dx}{d\sigma}\frac{\partial^2\sigma}{\partial n^2} > 0.$$

Proof of Theorem 2.2.3

Solving for $x(\phi)$, substituting, and proceeding as before,

$$u''\frac{dx}{d\phi} + \sigma f' + \sigma\phi^2 f''\frac{dx}{d\phi} + \sigma\alpha^2 g''\frac{dx}{d\phi} = 0.$$

$$dx/d\phi = \frac{-\sigma f'}{u'' + \sigma\phi^2 f'' + \sigma\alpha^2 g''} \gtreqless 0 \text{ if } f' \gtreqless 0 \text{ as } x \lesseqgtr Y - \eta.$$

Proof of Theorem 2.2.4

Substituting $x(\alpha)$ and differentiating,

$$u''\frac{dx}{d\alpha} + \sigma\phi^2 f''\frac{dx}{d\alpha} + \sigma g' + \sigma\alpha^2 g''\frac{dx}{d\alpha} = 0.$$

$$dx/d\alpha = \frac{-\sigma g'}{u'' + \sigma\phi^2 f'' + \sigma\alpha^2 g''} > 0.$$

Proof of Theorem 2.2.5

Substituting $x(\eta)$ and differentiating,

$$u''\frac{dx}{d\eta} + \sigma\phi^2 f''\frac{dx}{d\eta} - \sigma\phi^2 f'' + \sigma\alpha^2 g''\frac{dx}{d\eta} = 0.$$

$$0 < \frac{dx}{d\eta} = \frac{\sigma\phi f''}{u'' + \sigma\phi^2 f'' + \sigma\alpha^2 g''} < 1 \text{ if } \alpha \geq 0.$$

Proof of Theorem 3.1.3:

In the standard dictator game, the minimum and maximum transfers are zero and *X*, respectively.

Denote the null transfer x_N, where $x_N = 0$, and its frequency q_N. Denote the average super-fair transfer x_H, where $\frac{1}{2}X < x_H \leq X$, and its frequency q_H. Denote the average transfer between 0 and one-half x_G, where $0 < x_G \leq \frac{1}{2}X$, and its frequency q_G. Suppose $q_N, q_G, q_H \in (0,1)$. Finally, note that $q_N + q_G + q_H = 1$, and, according to SF 3.1.2, $q_H < q_N$. Then the average transfer equals

$$E(x) = q_N \cdot x_N + q_G \cdot x_G + q_H \cdot x_H = q_G \cdot x_G + q_H \cdot x_H.$$

First, note that $E(x) > 0$, since $x_N = 0$ and $q_G > 0, x_G > 0, q_H > 0$, and $x_H > 0$. Next, show $E(x) < \frac{1}{2}X$. Consider x_G at its maximum value $\frac{1}{2}X$, and x_H at its maximum value X. Note that $q_N \cdot x_N + q_H \cdot X = (q_N + q_H)\frac{q_H}{q_N+q_H} \cdot X < (1 - q_G)\frac{1}{2}X$, since $x_N = 0$, $q_N + q_H = 1 - q_G$, and $\frac{q_H}{q_N+q_H} < \frac{1}{2}$ from the fact that $q_H < q_N$. Then, the least upper bound of $E(x)$ is $\frac{1}{2}X$:

$$E(x) = q_G \cdot \frac{1}{2}X + [q_N \cdot x_N + q_H \cdot X] < q_G \cdot \frac{1}{2}X + (1 - q_G)\frac{1}{2}X = \frac{1}{2}X.$$

Proof of Theorem 3.1.5

$$U = u\big(\overline{M} - Y - x\big) + \sigma f\big(\phi(Y + x - \eta)\big) + \sigma\alpha g(\alpha x).$$

$$dU/_{dx} = -u'\big(\overline{M} - Y - x\big) + \sigma\phi f'\big(\phi(Y + x - \eta)\big) + \sigma\alpha g'(\alpha x) = 0.$$

Substituting $x(Y)$ and differentiating,

$$u'' + u''\frac{dx}{dY} + \sigma\phi^2 f'' + \sigma\phi^2 f''\frac{dx}{dY} + \sigma\alpha^2 g''\frac{dx}{dY} = 0,$$

$$-1 < \frac{dx}{dY} = \frac{-u'' - \sigma\phi^2 f''}{u'' + \sigma\phi^2 f'' + \sigma\alpha^2 g''} < 0.$$

Note that, in the absence of altruism, $\frac{dx}{dY} = -1$.

Proof of Theorem 3.1.6

Let *Y* denote the preset experimenter donation and *y* the amount by which the experimenter reduces the recipient's (R's) earnings. Then R earns $Y + x - y = Y$ since $x = y$.

$$U = u(X - x) + \sigma f\big(\phi(y - \eta)\big) + \sigma g(\alpha x),$$

$$dU/_{dx} = -u'(X - x) + \sigma\alpha g'(\alpha x) = 0$$

in the case of interior solutions. The assumptions that $A(\overline{\alpha} - \alpha^*) > 0$ and $\alpha^* \equiv \{\alpha | u'(X - \eta) = \sigma\alpha g'(\alpha\eta)\} > 0 \Longrightarrow u'(X) < \sigma\alpha g'(0) \ni \alpha^* > 0 \forall \alpha > \alpha^*$, who form the fraction $0.5 > A(\overline{\alpha} - \alpha^*) > 0$ plus some α for whom $\alpha^* \geq \alpha > 0$.

Proof of Theorem 3.1.7

Since the total stakes, $X + Y$, vary, the entitlement, η, would be impacted according to most distributive principles. Given the simple context, equality is a reasonable assumption, but I make the weaker assumption that $\eta = (1 - t)\overline{X} + tY$, $0 < t < 1$.

Then $Y + x - \eta = (1 - t)\big(Y - \overline{X}\big) + x$, and

$$U = u\big(\overline{X} - x\big) + \sigma f'\left(\phi\left((1 - t)\big(Y - \overline{X}\big) + x\right)\right) + \sigma g(\alpha x),$$

$$dU/_{dx} = -u'(\overline{X} - x) + \sigma\phi f''\left(\phi\left((1-t)(Y-\overline{X}) + x\right)\right) + \sigma\alpha g'(\alpha x) = 0.$$

Substituting $x(Y)$ and differentiating,

$$u''\frac{dx}{dY} + \sigma\phi^2(1-t)f'' + \sigma\phi^2 f''\frac{dx}{dY} + \sigma\alpha^2 g''\frac{dx}{dY} = 0,$$

$$-1 < \frac{dx}{dY} = \frac{-\sigma\phi^2(1-t)f''}{u''+\sigma\phi^2 f''+\sigma\alpha^2 g''} < 0.$$

Note that, in the absence of fairness, $\frac{dx}{dY} = 0$.

Proof of Theorem 4.1.1

$$dU/_{dz} = -u'(x + Z - z) + \sigma\phi f'\left(\phi\left(z - \frac{Z}{2} + \frac{X}{2} - x - \theta r(x-\tilde{x})\frac{Z}{2}\right)\right) + \sigma\alpha g'(\alpha x) = 0.$$

Substituting $z(x)$ and differentiating,

$$-u'' + u''\frac{dz}{dx} + \sigma\phi^2 f''\frac{dz}{dx} - \sigma\phi f'' - \sigma\phi^2\theta f'' r'\frac{Z}{2} + \sigma\alpha^2 g''\frac{dz}{dx} = 0,$$

$$dz/_{dx} = \frac{u''+\sigma\phi^2 f''+\sigma\phi^2\theta f'' r' Z/_2}{u''+\sigma\phi^2 f''+\sigma\alpha^2 g''}.$$

If x is randomly assigned, then $r = 0$ and $0 < \frac{dz}{dx} < 1$.

Otherwise, $r' > 0$ and sanctioning increases the value of $dz/_{dx}$.

Proof Theorem 4.1.2

$$dU/_{dz} = \sigma\phi f'\left(\phi\left(z - \frac{Z}{2} + \frac{X}{2} - x - \theta r(x-\tilde{x})\cdot\frac{Z}{2}\right)\right) = 0$$

$$\Longrightarrow z = \frac{Z}{z} - \frac{X}{z} + x + \theta r(x-\tilde{x})\frac{Z}{2}$$

$$dz/_{dx} = 1 + \theta r(x-\tilde{x})\frac{Z}{2}$$

which equals 1, if x is randomly assigned and $r = 0$, and is greater than 1, if x is chosen and $r' > 0$.

$d^2z/_{dx^2} = \theta r''\frac{Z}{2} < 0 \Longrightarrow$ asymmetric sanctioning.

Proof of Theorem 4.2.1

Since $\beta = 0$,

$$U = u(x_r) + \sigma f\left(\phi\left(z - \tfrac{1+b}{2}Z - \theta r(\hat{\gamma} - \tilde{\gamma})\cdot x'\right)\right).$$

$$dU/_{dz} = \sigma\phi f'\left(\phi\left(z - \frac{1+b}{2}Z - \theta\cdot r(\hat{\gamma} - \tilde{\gamma})\cdot x'\right)\right) = 0$$

$$\Longrightarrow z = \tfrac{1+b}{z}Z + \theta r(\hat{\gamma} - \tilde{\gamma})\cdot x'.$$

Distributive preferences imply the fixed amount $\frac{1+b}{z}Z$. The fact that $\hat{\gamma}^f > \hat{\gamma}^u$ implies the corresponding $z^f > z^u$ for a given value of x'.

Proof of Theorem 4.2.2

Let z^{x,x_r} be the R's choice of $\tilde{z}$ for the D's choice x and the realization x_r.

$z^{fF} = \frac{1+b}{z} Z + \theta r(\hat{\gamma}^f - \tilde{\gamma}) \cdot F$, since $x' = F$

$z^{uU} = \frac{1+b}{z} Z + \theta r(\hat{\gamma}^u - \tilde{\gamma}) \cdot F$, since $x' = F$

$z^{fU} = \frac{1+b}{z} Z + \theta r(\hat{\gamma}^f - \tilde{\gamma}) \cdot (qF + (1-q)L)$, since $x' = EX^f = qF + (1-q)L$

$z^{uF} = \frac{1+b}{z} Z + \theta r(\hat{\gamma}^u - \tilde{\gamma}) \cdot \big((1-q)F + qL\big)$ since $x' = EX^u = (1-q)F + qL$

$z^{fF} > z^{fU}$ as long as $z^{fU} < Z$ since $r > 0$ and $F > qF + (1-q)L$

$z^{uU} < z^{uF}$ since $r < 0$ and $F > (1-q)F + qL$

If D chooses with certainty, then for Rs

$$U = \sigma f(\phi\left(z - \frac{1+b}{2} Z - \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot F\right)$$

$$\Longrightarrow z = \tfrac{1+b}{2} Z + \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot F.$$

If D chooses fair,

$$z^F = \tfrac{1+b}{z} Z + \theta r(\hat{\gamma}^f - \tilde{\gamma}) \cdot F = z^{fF}.$$

If D chooses unfair,

$$z^U = \tfrac{1+b}{z} Z + \theta r(\hat{\gamma}^u - \tilde{\gamma}) \cdot F = z^{uU}.$$

Proof of Theorem 5.2.1

Define salience in the reference state to be at the level of a given first stage sinner and saints game, $\tilde{\sigma}$. Moreover, suppose, as usual, that $x^L > \underline{\gamma}$ and $x^H < \bar{\gamma}$. Then

$$\hat{\gamma}(x) = \begin{cases} \int_{\gamma^H}^{\bar{\gamma}} \gamma\rho(\gamma)d\gamma \Big/ \int_{\gamma^H}^{\bar{\gamma}} \rho(\gamma)d\gamma \;\; if\; x = x^H \\ \qquad x(\tilde{\sigma})\; if\; x^L < x < \; x^H \\ \int_{\underline{\gamma}}^{\gamma_L} \gamma\rho(y)d\gamma \Big/ \int_{\underline{\gamma}}^{\gamma_L} \rho(y)d\gamma \;\; if\; x = x^L \end{cases}$$

The first order condition is

$$dU/_{dz} = \sigma\phi f'\left(\phi\left(z - \eta_z - \theta r(\hat{\gamma} - \tilde{\gamma}), \frac{Z}{2}\right)\right) = 0$$

$$\Rightarrow z = \frac{Z}{2} - \frac{X}{2} - \frac{Y}{2} + x + \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot \frac{Z}{2}$$

When $x^L < x < x^H$, $\hat{\gamma} = x$, and

$dz/_{dx} = 1 + \theta r'(\hat{x} - \tilde{x})\ \frac{Z}{2}$, and

$d^2z/_{dx^2} = \theta r'' \frac{Z}{2} < 0.$

When $x = x^L$,

$$\tilde{z}\Big|_{x^L} = \frac{Z}{2} - \frac{X}{2} - \frac{Y}{2} + x + \theta r(\hat{\gamma}(x^L) - \tilde{\gamma}) \frac{Z}{2}$$

$$< \tilde{z}\Big|_{\gamma^L} = \frac{Z}{2} - \frac{X}{2} - \frac{Y}{2} + x + \theta r\ (\gamma^L - \tilde{\gamma})\ \frac{Z}{2}$$

since $\tilde{\gamma}(x^L) < \gamma^L$.

When $x = x^H$,

$$\tilde{z}\Big|_{x^H} = \frac{Z}{2} - \frac{X}{2} - \frac{Y}{2} + x + \theta r(\hat{\gamma}(x^H) - \tilde{\gamma})\frac{Z}{2}$$
$$> \tilde{z}\Big|_{\gamma^H} = \frac{Z}{2} - \frac{X}{2} - \frac{Y}{2} + x + \theta r(\gamma^H - \tilde{\gamma})\frac{Z}{2}$$

since $\hat{\gamma}(x^H) > \gamma^H$. This implies $\tilde{z}$ is concave in x with a discontinuous decrease at x^L and a discontinuous increase at x^H. By Assumption 7, $dn/dx^L < 0$, by Definition 1, $\partial\sigma/\partial n < 0$, and by Theorem 2.2.3, $dx/d\sigma > 0$. Then if a D gives $\tilde{\gamma}$ at the reference salience $\tilde{\sigma}$, then increasing x^L increases the threshold transfer in the sinners and saints game:

$$\partial\tilde{x}/\partial x^L = \frac{dx}{d\sigma}\frac{\partial\sigma}{\partial n}\frac{dn}{x^L} > 0$$

for interior solutions. Note that $x(\tilde{\gamma}, \sigma)$, so if x is held constant,

$$dx = \frac{dx}{d\hat{\gamma}}d\hat{\gamma} + \frac{\partial x}{\partial\sigma}d\sigma = 0$$
$$\Rightarrow \frac{d\hat{\gamma}}{d\sigma}\Big|_x = \frac{\partial x/\partial\sigma}{\partial x/\partial\hat{\gamma}} = -\frac{\partial x}{\partial\sigma} < 0 \text{ since } \frac{\partial x}{\partial\hat{\gamma}} = 1$$

for interior solutions. Then

$$\frac{\partial\hat{\gamma}}{\partial x^L}\Big|_x = \frac{d\hat{\gamma}}{d\sigma}\frac{\partial\sigma}{\partial n}\frac{dn}{dx^L} < 0$$

and $\frac{\partial z}{\partial x^L} = \frac{\partial z}{\partial\hat{\gamma}}\frac{\partial\hat{\gamma}}{\partial x^L} < 0$ since $\frac{\partial z}{\partial\hat{\gamma}} > 0$.

Proof of Theorem 5.2.2

$$U = u(\bar{z}) + \sigma f(\phi\left(z - \frac{Z}{2} + \eta_x(x^L) - x - \theta r\big(x - \tilde{x}(x^L)\big)\cdot\frac{Z}{2}\right))$$
$$dU/dz = \sigma\phi f'\left(\phi\left(z - \frac{Z}{2} + \eta_x(x^L) - x - \theta r\big(x - \tilde{x}(x^L)\big)\cdot\frac{Z}{2}\right)\right) = 0$$
$$\Rightarrow z = \frac{Z}{2} - \eta_x(x^L) + x + \theta r\big(x - \tilde{x}(x^L)\big)\cdot\frac{Z}{2}))$$
$$\frac{dz}{dx^L}\Big|_x = -\frac{d\eta_x}{dx^L} - \theta r'\big(x - \tilde{x}(x^L)\big)\cdot\frac{Z}{2}\cdot\frac{d\tilde{x}}{dx^L} < 0$$

Proof of Theorem 5.3.1

When $x = Y = M, \eta = \frac{M}{2}$, and

$$U = u(X) + \sigma f(\phi x) + \sigma g(\alpha x).$$
$$dU/dx = \sigma\phi f'(\phi x) + \sigma\alpha g'(\alpha x) = 0$$
$$\Rightarrow \sigma\alpha g'(\alpha x) = -\sigma\phi f'(\phi x)$$
$$\Rightarrow \alpha \gtreqless 0 \Rightarrow f' \lesseqgtr 0 \Rightarrow x \gtreqless 0.$$

If x is constrained, $x \leq 0$, then $x = 0$ for $\alpha \geqq 0$ and $x < 0$ only for $\alpha < 0$, which is a minority according to the assumption that $0 < \mathrm{A}(0) < 0.5$. When $X > Y$, the agent's utility function is

$$U = u(X) + \sigma_f f\left(\phi\big(x - Y - \mu/2\big)\right) + \sigma_g g(\alpha x)$$

allowing separate variation in σ_f and σ_g, where I assume $0 \leq \frac{d\sigma_f}{d\sigma_g} = \kappa \leq 1$.

$$dU/_{dx} = \sigma_f \phi f'\left(\phi\left(x + Y - M/_2\right)\right) + \sigma_g \alpha g'(\alpha x) = 0$$

Solving $x(\sigma_g)$ and differentiating with respect to σ_g

$$\kappa\phi f' + \sigma_f \phi^2 f'' \frac{dx}{d\sigma_g} + \alpha g' + \sigma_g \alpha^2 g'' \frac{dx}{d\sigma_g} = 0$$

$$dx/_{d\sigma_g} = \frac{-\kappa\phi f' - \alpha g'}{\sigma_f \phi^2 f" + \sigma_g \alpha^2 g"}$$

If $\sigma_f = \sigma_g = \sigma$, then $\kappa = 1$, and $\frac{dx}{d\sigma_g} = 0$ from the first order condition. Suppose $\alpha > 0$, and initially $\sigma_f = \sigma_g$ and $\kappa < 1$, or $\sigma_f \neq \sigma_g$ but $\kappa = 0$. Then $dx/_{d\sigma_g} > 0$. If $\alpha < 0$, then altruism salience is $\sigma_b = 1 - \sigma_g$, $dx/_{d\sigma_g} < 0$, and

$$dx/_{d\sigma_b} = \frac{dx}{d\sigma_g}\frac{d\sigma_g}{d\sigma_b} = -\frac{dx}{d\sigma_g} > 0.$$

Proof of Theorem 5.3.2

If endowments are fair, then $Y = \eta$, and Theorem 5.3.1 applies, including to cases when $\eta \neq \frac{M}{2}$. If a patient is unfairly advantaged, i.e., $Y > \eta$, whether with earned or unearned endowments, then, by Theorem 5.3.1 while simplifying to a single moral salience term,

$$\sigma\alpha g'(\alpha x) = -\sigma\phi f'(\phi(Y + x - \eta))$$

All spiteful agents destroy, since $\alpha < 0 \Rightarrow f' > 0 \Rightarrow x < \eta - Y < 0$. Also, $\alpha = 0$ destroy: $f' = 0 \Rightarrow x = \eta - Y$. The critical α for destroying, α^D, solves $x = 0$ for

$$\sigma\alpha g'(0) = -\sigma\phi f\left(\phi(Y - \eta)\right) > 0$$

$$\Rightarrow \alpha^D = \frac{-\sigma\phi f'(\phi(Y - \eta))}{\sigma g'(0)} > 0$$

which implies even some altruistic agents destroy.

Proof of Theorem 6.1.1

A dictator enters if

$$UN = u(X - x) + \sigma f\left(\phi(x - \eta)\right) + \sigma g(\alpha x) > UX = (X - c)$$

Of those who enter, more self-interested Ds transfer nothing: at a minimum, those with $\alpha = 0$ and $\phi = \underline{\phi}$ by Theorem 3.1.1. Some who enter make positive transfers: at a minimum, super-fair Ds by Theorem 3.1.2. From Theorem 2.2.3, $dx/_{d\phi} > 0$ and from Theorem 2.2.4 $dx/_{d\alpha} > 0$. Then

$$dUN/_{d\phi} = -u'\frac{dx}{d\phi} + \sigma(x - \eta)f' + \sigma\phi f'\frac{dx}{d\phi} + \sigma\alpha g'\frac{dx}{d\phi}$$

$$= (-u' + \sigma\phi f' + \sigma a g')\frac{dx}{d\phi} = \sigma(x - \eta)f'$$

$$= \sigma(x - \eta)f' < 0$$

for the mean case and since $-u' + \sigma\phi f' + \sigma\alpha g' = 0$ for an interior solution and $dx/d\phi = 0$ for a corner solution.

$$dUN/d\alpha = (-u' + \sigma\phi f' + \sigma\alpha g')\frac{dx}{d\alpha} + \sigma x g'$$
$$= \sigma\alpha g' > 0 \text{ since } -u' + \sigma\phi f' + \sigma\alpha g' = 0$$

for an interior solution and $dx/d\alpha = 0$ for a corner solution. If altruism is second order, then fairer Ds exist since the utility of entering is decreasing in ϕ. The threshold ϕ for exiting is

$$UX > UN$$
$$u(X-c) > u(X-x) + \sigma\phi^{\lambda} f(x-\eta) + \sigma g(\alpha x)$$
$$\phi^{\lambda}\sigma f(x-\eta) < u(X-c) - u(X-x) - \sigma g(\alpha x)$$
$$\phi > \phi^{x} = \left[\frac{u(X-c) - u(X-x) - \sigma g(\alpha x)}{\sigma f(x-\eta)}\right]^{1/\lambda}$$

Proof of Theorem 6.1.3

UX is fixed at $u(\bar{X} - c)$, since $\bar{X}$ and c are fixed. The first order condition for the mean D with an interior solution who enters is

$$dUN/dx = -u'(X-x) + \sigma\phi f'\big(\phi(x-\eta)\big) + \sigma\alpha g'(\alpha x) = 0$$

Assume quite generally that $0 \le \frac{d\eta}{dX} \le 1$. Differentiating with respect to $X, x(X)$ and $\eta(X)$,

$$-u'' + u''\frac{dx}{dX} + \sigma\phi^2 f''\frac{dx}{dX} - \sigma\phi^2 f''\frac{d\eta}{dX} + \sigma\alpha^2 g''\frac{dx}{dX} = 0$$
$$\Rightarrow 0 < \frac{dx}{dX} = \frac{u" + \sigma\phi^2 f''\, d\eta/dX}{u'' + \sigma\phi^2 f'' + \sigma\alpha^2 g''} < 1$$

The effect of X on UN evaluated at the optimal x is

$$\frac{dUN}{dX}\Big|_x = u' - u'\frac{dx}{dX} + \sigma\phi f'\frac{dx}{dX} - \sigma\phi f'\frac{d\eta}{dX} + \sigma\alpha g'\frac{dx}{dX}$$
$$= u' - \sigma\phi f'\frac{d\eta}{dX} + (-u' + \sigma\phi f' + \sigma\alpha g')\frac{dx}{dX}$$
$$= u' - \sigma\phi f'\frac{d\eta}{dX} > 0$$

since, by the first order condition for the mean D, $u' > \sigma\phi f'$ and since $0 \le d\eta/dX < 1$. That is, as X rises, so does UN, and more Ds choose entry.

Proof of Theorem 6.2.1

To simplify notation, let salience under reveal be $\sigma^m = 1$ and under not reveal $0 < \sigma^l \equiv \sigma < 1$.

$R2B$ is dominated

$$U(R2A) > U(R2B)$$
$$u(H) + f\big(\phi(F-F)\big) + g(\alpha F) > u(F) + f\big(\phi(L-F)\big) + g(\alpha L)$$

since all LHS terms $>$ all RHS terms for the mean and most Ds. For a minority $\alpha < 0$, which could reverse the inequality if D were extremely spiteful, but we assume u and f dominate g.

$\underline{NRB}$ is dominated

$$EU(R1B, R2A) = .5U(R1B) + .5U(R2A) > EU(NRB)$$
$$.5u(F) + .5g(\alpha F) + .5u(H) + .5g(\alpha F)$$
$$> u(F) + .5\sigma g(\alpha F) + .5\sigma f\big(\phi(L-F)\big) + .5\sigma g(\alpha L)$$
$$u(H) - u(F) - \sigma f\big(\phi(L-F)\big) + (2-\sigma)g(\alpha F) + \sigma g(\alpha L) > 0$$

if $\alpha \geq 0$. This inequality could only be reversed, if D were very spiteful but less than for $R2B$. Note also that NRB is dominated by NRA, since $EU(NRA) - EU(NRB) = u(H) - u(F) > 0$.

Fairer Ds choose R1A and R2A over NRA

$$EU(R1B, R2A) = .5U(R1B) + .5U(R2A) > EU(NRA)$$
$$.5u(F) + .5g(\alpha F) + .5u(H) + .5g(\alpha F)$$
$$> u(H) + .5\sigma f\big(\phi(L-F)\big) + .5\sigma g(\alpha L) + .5\sigma g(\alpha F)$$
$$\Rightarrow \phi > \phi^{NRA} = \left[\frac{\frac{1}{\sigma}[u(F) - u(H)] + \left(\frac{2-\sigma}{\sigma}\right)g(\alpha F) - g(\alpha L)}{f(L-F)}\right]^{1/\lambda} > 0$$

since g is second order.

Least fair Ds choose $R1A$ and $R2A$ over NRA

$$EU(R1A, R2A) = .5U(R1A) + .5U(R2A) > EU(NRA)$$
$$.5u(H) + .5f(\phi(L-F)) + .5g(\alpha L) + .5u(H) + .5g(\alpha F)$$
$$> u(H) + .5\sigma f(\phi(L-F)) + .5\sigma g(\alpha L) + .5\sigma g(\alpha F)$$
$$\Rightarrow \phi > \phi^{R12A} \equiv \left[-\frac{g(\alpha L) + g(\alpha F)}{f(L-F)}\right]^{1/\lambda} \geq 0$$

for $\alpha > 0$ and corner solution at zero for $\alpha \leq 0$. Note $\phi^{R12A} < \phi^{NRA}$ since g is second order. In light of dominated strategies, the choices reduce to $R1A2A$, NRA and $R1B2A$. Letting fairness types be denoted, in ascending order of strength of fairness preferences, we have the following choices ordered by fairness type:

$$R1A2A < \phi^{R12A} < NRA < \phi^{NRA} < R1B2A.$$

Standard game versus information game

In the standard game, salience is the highest, σ^h, and D chooses $1B$ over $1A$ if

$$u(1B) > u(1A)$$
$$u(F) + \sigma^h f(\phi(F-F)) + \sigma^h g(\alpha F) > u(H) + \sigma^h f(\phi(L-F)) + \sigma^h g(\alpha L)$$
$$\Rightarrow \phi > \phi^{1B} \equiv \left[\frac{\frac{1}{\sigma^h}[u(F) - u(H)] + g(\alpha F) - g(\alpha L)}{f(L-F)}\right]^{1/\lambda}.$$

Assuming g is second order, $\phi^{B1} > \phi^{R12A}$, and also noting $\sigma^h > \sigma^m = \sigma$, $\phi^{B1} < \phi^{NRA}$. Thus, the fraction choosing $1B$ in the standard game should be greater than the fraction choosing $R1B2A$ and less than the fraction choosing $R1B2A$ or NRA in the information game. DWK did not run a baseline for $2A$ versus $2B$, but BEW did, and we will use the following result in the next proof. In the standard game, $u(1B) > u(1A)$ if

$$u(H) + \sigma^h f(\phi(F-F)) + \sigma^h g(\alpha F) > u(F) + \sigma^h f(\phi(L-F)) + \sigma^h g(\alpha L)$$

$$\Rightarrow \phi > \phi^{2A} \equiv \left[\frac{\frac{1}{\sigma^h}[u(H) - u(F)] + g(\alpha F) - g(\alpha L)}{f(L - F)}\right]^{1/\lambda}.$$

Note $\phi^{2A} \geq 0$ only if D is extremely spiteful.

Proof of Theorem 6.2.2

Fairness and altruism are assumed not to be negatively correlated, so generosity in the reference state, γ, is positively correlated with ϕ and α, according to Theorems 2.2.3 and 2.2.4, respectively. The possible exception is super-fair Ds, but they are ruled out by design in the information game. Thus, γ is replaced by ϕ and α in the following proofs, since only order matters with respect to γ. From the proof for Theorem 6.2.1, one can conclude that the estimated ϕ for each choice is, in descending order, $R1B(\hat{\phi}^{R1B})$, $1B(\hat{\phi}^{1B})$, $NRA(\hat{\phi}^{NRA})$, $R1A(\hat{\phi}^{R1A})$, and $1A(\hat{\phi}^{1A})$. Also, $R2A(\hat{\phi}^{R2A})$ lies somewhere between $R1B$ and $R1A$, since it consists of a mix of these two types and is less than $2A(\hat{\phi}^{2A})$. Two other choices are predicted to be dominated, but, if they occurred, would indicate very spiteful types, where $R2B(\hat{\alpha}^{R2B})$ is more spiteful than $NRB(\hat{\alpha}^{NRB})$, but $2B(\hat{\alpha}^{2B})$ is more spiteful than both. The thresholds for ϕ and α are assumed to be, respectively, between $1B$ and NRA ($\hat{\phi}^{1B} > \tilde{\phi} > \hat{\phi}^{NRA}$) and above NRB ($\tilde{\alpha} > \hat{\alpha}^{NRB}$). The utility function of a spectator is

$$U = u(\bar{z}) + \sigma(f(\phi(z - \eta_z - \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot x')$$

where $\eta_z = 0$ since $\beta = 0$ and $z \leq 0$.

$$\frac{dU}{dZ} = \sigma\phi f'(\phi(z - \theta r(\hat{\gamma} - \tilde{\gamma})) \cdot x') = 0$$

$$\Rightarrow \tilde{z} = \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot x'$$

The remaining analysis is based on replacing $\hat{\gamma}$ with $\hat{\phi}$ and $\hat{\alpha}$ to save the notation $\hat{\gamma}(\hat{\phi})$ and $\hat{\gamma}(\hat{\alpha})$, and similarly for $\tilde{\phi}$ and $\tilde{\alpha}$. For $R1B$, $\hat{\phi}^{R1B} > \tilde{\phi}$, and for $1B$, $\hat{\phi}^{1B} > \tilde{\phi} \Rightarrow \tilde{z} > 0$, but z is constrained to $z \leq 0$, so $z = 0$.

$$z^{NRA} = \theta r\left(\hat{\phi}^{NRA} - \tilde{\phi}\right) \cdot x' < 0$$

NRA is an unfair choice, since $R1B$ was possible and $\hat{\phi}^{NRA} < \tilde{\phi}$, so if the unfair state 1 obtains, $x' = F$. If the fair state 2 obtains, $x' = EX^u = .5F + .5L = EX^f \equiv EX < F$ in this game. Thus, the punishments based on realizations are

$$z^{NRA1} = \theta r\left(\hat{\phi}^{NRA} - \tilde{\phi}\right) \cdot F$$

$$z^{NRA2} = \theta r\left(\hat{\phi}^{NRA} - \tilde{\phi}\right) \cdot EX$$

For the remaining choices involving $\hat{\phi}$, $\tilde{z}$ is

$$z^{R1A} = \theta r\left(\hat{\phi}^{R1A} - \tilde{\phi}\right) \cdot F$$

$$z^{1A} = \theta r\left(\hat{\phi}^{1A} - \tilde{\phi}\right) \cdot F < z^{2A} = \theta r\left(\hat{\phi}^{2A} - \tilde{\phi}\right) \cdot F$$

$$z^{R2A} = \theta r\left(\hat{\phi}^{R2A} - \tilde{\phi}\right) \cdot F$$

Similarly,

$$z^{NRB1} = \theta r(\hat{\alpha}^{NRB} - \tilde{\alpha}) \cdot EX$$

$$z^{NRB2} = \theta r(\hat{\alpha}^{NRB} - \tilde{\alpha}) \cdot F$$

$$z^{R2B} = \theta r(\hat{\alpha}^{R2B} - \tilde{\alpha}) \cdot F \quad < z^{2B} = \theta r(\hat{\alpha}^{2B} - \tilde{\alpha}) \cdot F$$

These imply the following

$$z^{R1B} = z^{1B} = 0 > z^{NRA2} > z^{NRA1} > z^{R1A} > z^{1A}$$
$$z^{R1B} > z^{2A} > z^{R2A} > z^{R1A}$$
$$0 > z^{NRB1} > z^{NRB2} > z^{R2B} > z^{2B}$$

Proof of Theorem 6.3.1

In the baseline, salience is high, σ^h, and the D chooses fair if

$U(F) > U(U)$

$$u(F) + \sigma^h 3g(\alpha F) > u(H) + \sigma^h f\big(\phi(H-F)\big) + \sigma^h 2f\big(\phi(L-F)\big) + \sigma^h g(\alpha H) + \sigma^h 2g(\alpha L)$$

$$\phi > \phi^B \equiv \left[\frac{\frac{1}{\sigma^h}[u(F)-u(H)] - [g(\alpha H) + 2g(\alpha L)] + 3g(\alpha F)}{f(H-F) + 2f(L-F)}\right]^{1/\lambda}$$

In the delegation game, the D chooses fair with salience σ^m over delegating with salience σ^l if

$U(F) > EU(D)$

$$u(F) + \sigma^m 3g(\alpha F) > qu(F) + (1-q)u(H) + (1-q)\sigma^l f\big(\phi(H-F)\big) + (1-q)\sigma^l 2f\big(\phi(L-F)\big) + q\sigma^l 3g(\alpha F) + (1-q)\sigma^l g(\alpha H) + (1-q)\sigma^l 2g(\alpha L)$$

$$\phi > \phi^F \equiv \left[\frac{\frac{1}{\sigma^l}[u(F)-u(H)] - [g(\alpha H) + 2g(\alpha L)] + \frac{\sigma^m - q\sigma^l}{\sigma^l(1-q)} \cdot 3g(\alpha F)}{f(H-F) + 2f(L-F)}\right]^{1/\lambda}$$

In the delegation game, the D chooses unfair over delegating if

$U(U) > EU(D)$

$$u(H) + \sigma^m f\big(\phi(H-F)\big) + \sigma^m 2f\big(\phi(L-F)\big) + \sigma^m g(\alpha H) + \sigma^m 2g(\alpha L) > qu(F) + (1-q)u(H) + (1-q)\sigma^l f\big(\phi(H-F)\big) + (1-q)\sigma^l 2f\big(\phi(L-F)\big) + q\sigma^l 3g(\alpha F) + (1-q)\sigma^l g(\alpha H) + (1-q)\sigma^l 2g(\alpha L)$$

$$\phi < \phi^D \equiv \left[\frac{\frac{1}{\sigma^l + \frac{1}{q}(\sigma^m - \sigma^l)}[u(F)-u(H)] - [g(\alpha H) + 2g(\alpha L)] + \frac{q\sigma^l}{\sigma^m - (1-q)\sigma^l} \cdot 3g(\alpha F)}{f(H-F) + 2f(L-F)}\right]^{1/\lambda}$$

Since altruism is assumed second order, we focus on the other terms. Then, Ds with $\phi > \phi^F$ allocate fairly. Note $\phi^F > \phi^D$ since $\sigma^l < \sigma^l + \frac{1}{q}(\sigma^m - \sigma^l)$, so Ds with $\phi^F > \phi > \phi^D$ delegate. Then those with $\phi < \phi^D$ allocate unfairly. $\phi^F > \phi^B$ since $\sigma^l < \sigma^h$, so fewer Ds allocated fairly in the delegation game than in the dictator game. Note the fraction allocating unfairly in the two games is ambiguous since

$\phi^B \gtreqless \phi^D$ as $\sigma^h \lesseqgtr \sigma^l + \frac{1}{q}(\sigma^m - \sigma^l)$.

The intermediary chooses fair if

$\underline{U(F) > U(U)}$

$$u(F) + \sigma^I 3g(\alpha F) > u(H) + \sigma^I f\big(\phi(H-F)\big) + \sigma^I 2f\big(\phi(L-F)\big) + \sigma^I g(\alpha H) + \sigma^I 2g(\alpha L)$$

$$\phi > \phi^I \equiv \left[\frac{\frac{1}{\sigma^I}[u(F)-u(H)] - [g(\alpha H) + 2g(\alpha L)] + 3g(\alpha F)}{f(H-F) + 2f(L-F)}\right]^{1/\lambda}$$

Disregarding altruism terms, $\sigma^I = \sigma^l \Rightarrow \phi^I = \phi^F$, and $\sigma^I = \sigma^m \Rightarrow \phi^I > \phi^D$ since

$$\sigma^m > \sigma^l + \frac{1}{q}(\sigma^m - \sigma^l),$$

hence, $\phi^F \geq \phi^I > \phi^D$, and the fraction of Is allocating fairly are, therefore, those with ϕ greater than this range of values.

Proof of Theorem 6.3.2

The spectator's utility function is

$$U = u(\bar{z}) + \sigma f(\phi(z - \theta r(\hat{\phi} - \tilde{\phi}) \cdot x',$$

replacing γ with ϕ, as in section 6.2, where altruism is second order,

$$\Rightarrow \tilde{z} = \theta r\big(\hat{\phi} - \tilde{\phi}\big) \cdot x'.$$

Whenever fair is chosen, $\hat{\phi} > \tilde{\phi}$, and reward is optimal, but this is ruled out by $z \leq 0$. Those, who are in passive roles, cannot reveal their character, so $z = 0$ for those cases, as well. That leaves the following cases. In the baseline, the dictator who chooses unfair has an estimated $\hat{\phi}^B, \phi^B > \hat{\phi}^B > \underline{\phi}$, where $\hat{\phi}^B < \tilde{\phi}$. The optimal sanction solves

$$z_D^{BU} = \theta r\big(\hat{\phi}^B - \tilde{\phi}\big) \cdot F < 0$$

In the delegation game, the D who chooses unfair has $\phi^D > \hat{\phi}^U > \underline{\phi}$ and the optimal sanction is

$$z_D^{DU} = \theta r\big(\hat{\phi}^U - \tilde{\phi}\big) \cdot F < 0$$

Note $z_D^{BU} \gtreqless z_D^{DU}$ as $\phi^B \lesseqgtr \phi^D$. The I who chooses unfair has $\phi^I > \hat{\phi}^I > \underline{\phi}$ and the sanction is

$$z_I^{DDU} = \theta r\big(\hat{\phi}^I - \tilde{\phi}\big) \cdot F < 0$$

Note $z_I^{DDU} < z_D^{DDU}$ depending on $\phi^I > \phi^D$. Finally, if the threshold is above $\hat{\phi}^D$, then the D is also punished for delegating. Note $\hat{\phi}^D < \tilde{\phi} \leq \phi^F$ and

$$z_D^{DD} = \theta r\big(\hat{\phi}^D - \tilde{\phi}\big) \cdot x'$$

where if unfair obtains, $x' = EX^u$, and if fair obtains, $x' = EX^f$, where $q < 0.5$, so $EX^u > EX^f$, resulting in $0 > z_D^{DDF} > z_D^{DDU}$, if $\tilde{\phi} < \phi^F$, and $0 = z_D^{DDF} = z_D^{DDU}$, otherwise.